

REPORT STI-TR-948-02-F

Scientific and Technical Report – Final Report
Contract DAAE07-02-C-L016: CDRL CLIN/ELIN 0002/A002
An Integrated Suite of Text and Data Mining Tools – Phase II

Dated: August 30, 2005

Contract Period: May 22, 2002 through August 30, 2005

Submitted (via e-mail) to:

U.S. Army Tank-automotive and Armaments Command
AMSTA-TR-N/272 (Mr. Bob Watts)
Warren, MI 48397-5000

U.S. Army Tank-automotive and Armaments Command
AMSTA-AQ-ABGB, M/S 321 (Mr. Doug Schroeder)
Warren, MI 48397-5000

Defense Technical Information Center (DTIC)
Cameron Station, VA

Prepared by:

Paul R. Frey, Brian S. Minsk, and Alan L. Porter
Search Technology, Inc.
4960 Peachtree Industrial Blvd.
Suite 230
Norcross, GA 30071

Report Documentation Page				Form Approved OMB No. 0704-0188	
Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.					
1. REPORT DATE 00 MAR 2005		2. REPORT TYPE N/A		3. DATES COVERED -	
4. TITLE AND SUBTITLE An Integrated Suite of Text and Data Mining Tools, Phase II,				5a. CONTRACT NUMBER DAAE07-02-C-L016	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S)				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Search Technology, Inc., 4960 Peachtree Industrial Blvd., STE 230, Norcross, GA 30071				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release, distribution unlimited					
13. SUPPLEMENTARY NOTES This distribution was called on and "public release" is what they want per Kathy Velez (770) 441-1458, Ext 156., The original document contains color images.					
14. ABSTRACT					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT UU	18. NUMBER OF PAGES 67	19a. NAME OF RESPONSIBLE PERSON
a. REPORT unclassified	b. ABSTRACT unclassified	c. THIS PAGE unclassified			

Abstract

This report summarizes the results of a three-year SBIR project to develop an integrated suite of text and data mining tools. The goal of this project is to provide tools that can help analysts find connections between requirements (as expressed in requirements documents or databases) and open-source research literature. An overall approach is outlined, and a step-by-step overview of the work is presented. The tool suite includes parsers for text data sources, metadata extraction, record combining, entity extraction, data normalization, sub- and cross-dataset analysis, multi-field analysis and visualizations, feature selection, XML importers, and indirect link analysis. A set of recommendations for expanding the use of the tools is presented.

Abstract	ii
Summary	1
Introduction	2
Methods, Assumptions, and Procedures	3
Results and Discussion	5
Segment the Requirements Document	6
Extract Metadata	8
Parse the Segments	8
Extract Entities	9
Reduce and Combine Data	11
Decompose Segments	12
Mine the Source Document	12
Refine the Query	13
Mine Open Literature Dataset	15
Automation/Macros	20
Leverage Open-Source Data Repositories	22
Quick XML Import	23
Simultaneous Cross Dataset Analysis	24
Update Datasets	25
Graphical Display of Numeric Data	27
WebQL interface	28
Link Analysis	28
Indirect Link Matrix	29
Matrix List	30
Link Analysis Enhancements to the Detail Window	32
Link Analysis Example	33
Conclusions	38
Recommendations	41
References	43
Appendix A	44
Appendix B	56

Summary

The goal of this project is to provide a suite of text- and data-mining tools that can help analysts find connections between requirements (as expressed in requirements documents or databases) and open-source research literature. The purpose in making these connections is to help the analyst identify R&D results in the open literature that augment, support, or clarify the requirements, and identify and qualify outside capabilities and competencies that may be employed to help satisfy the requirement.

This project builds upon an ongoing line of research and development activities at Search Technology. These R&D activities grew from work at Georgia Tech in the 1980s and 1990s. Beginning in 1995 under a DARPA STTR project (eventually carried through 3 phases), Search Technology in collaboration with Georgia Tech Research Corporation developed a desktop text-mining software tool named TechOASIS (known commercially as VantagePoint). By the end of that STTR project, TechOASIS (or VantagePoint) was a viable product in both government and commercial markets, providing technology managers and competitive intelligence professionals a tool for analyzing bibliographic search results of R&D and patent literature, supporting technology assessment, monitoring, and forecasting.

The current project seeks to radically expand the functionality of TechOASIS beyond bibliographic data sources to less-structured text sources, and to expand the application to leveraging technology assessment and forecasting against requirements analysis.

The project met its major objectives. This report describes over a dozen major advances to the tool suite and many more minor enhancements, all of which have been made available to government users. The enhancements include:

- Fully revamped import engine based on Regular Expressions
- Import Engine Editor
- Import Variables
- Frankenrecords
- Entity Extraction
- “Protected” NLP parsing
- Extracting phrases in context (“nearby phrases”)
- Multi-stage “find” with Boolean and proximity searching
- Save/resume data normalization (list cleanup)
- Detail Windows
- Expectancy arrows in Detail Windows
- Cross-field displays on maps (categorical data; graphical pop-ups)
- Analysis using subsets of the dataset
- TFIDF matrix
- Quick XML import
- Cross-dataset analysis (virtual fields/virtual datasets)
- Indirect Link matrix

- Matrix List

Most of these advances have already been incorporated into two commercial products, continuing a success story for the SBIR/STTR program.

Introduction

The main objective of Phase II is to provide a suite of tools that can help a user:

- Discover clusters of relationships in a requirements document or database and
- Find relationships between those clusters and open-source research literature,

with the ultimate goals of:

- Identifying R&D results in the open literature that augment, support, or clarify the user's requirements, and
- Identifying and qualifying outside capabilities and competencies that may be employed to help accomplish the user's objectives and goals.

During the course of Phase II, we addressed the following technical objectives:

1. Techniques for text mining and data mining – For our target user population and their data sources and tasks, what are the best methods and tools for mining knowledge from text and data?
2. Parsing techniques for free text – For the data sources and the text- and data-mining techniques, what are the best methods for parsing words/phrases from the free text portions of the data sources?
3. Information Extraction – For the user's tasks and data sources, what method will work best for extracting user-specified information from free text?
4. Information Reduction – What thesauri and other techniques should be used to combine similar word/phrases and data to result in a more usable representation of the actual content of the data?
5. Combining text mining and data mining to find, associate, and evaluate relational groupings – How should we combine the techniques of text- and data-mining to produce an integrated, cohesive suite of tools? One of the primary objectives of the tool suite is to *discover relational groupings* in text/data, *discover associations* among those groupings, and *quantify the cohesiveness of the groupings* and the *strength of the associations*.
6. Cross-mining requirements and bibliographic databases – How can we provide our users with the maximum benefit from mining their requirements databases *and* bibliographic databases of the open S&T literature? Knowledge discovered in one data source can provide insight into how to mine the other data source.
7. Hosting/Architecture – For the user population, their data sources, and our system design, what architecture should be used? We began Phase II with a text-mining

tool that operates on stand-alone personal computers. Do other host/architecture solutions that make sense for integrating existing data- and text-mining tools in this project?

During the course of Phase II, other organizations invested in the project via Phase II Plus to advance particular objectives.

1. The National Institutes of Health/National Library of Medicine funded a project pioneering the application of text- and data-mining tools to their Hazardous Substances Database (HSDB). A matching Phase II Plus award funded the investigation of using Internet-based sources of open-source R&D information and comparing the effectiveness of those sources with fee-based sources.
2. Sandia National Laboratory funded a project aimed at incorporating the text- and data-mining tools developed under Phase II into a broader suite of tools named Advanced Decision Support System (or ADSS). Another matching Phase II Plus award funded several initiatives including simultaneous cross-dataset analysis, automated import of XML documents, and link analysis, among others.

Methods, Assumptions, and Procedures

The text mining tool suite can be divided into two main activities – 1) mining knowledge from requirements documents and 2) using knowledge mined from requirements documents to mine open research and development literature. Each of these activities is first illustrated with a diagram and each step in the process (each circle) is described in a subsequent paragraph.

Figure 1 illustrates the activity of mining knowledge from requirements documents. It begins with slicing the source document into basic analysis “chunks” or “segments” and continues with extracting data from the segments and reducing the data to a minimal set of symbols (e.g., root words or numeric categories) representing unique meanings. Then the collection of segments is decomposed (or partitioned), and finally knowledge is mined from the source document using information extracted and derived from the segments.

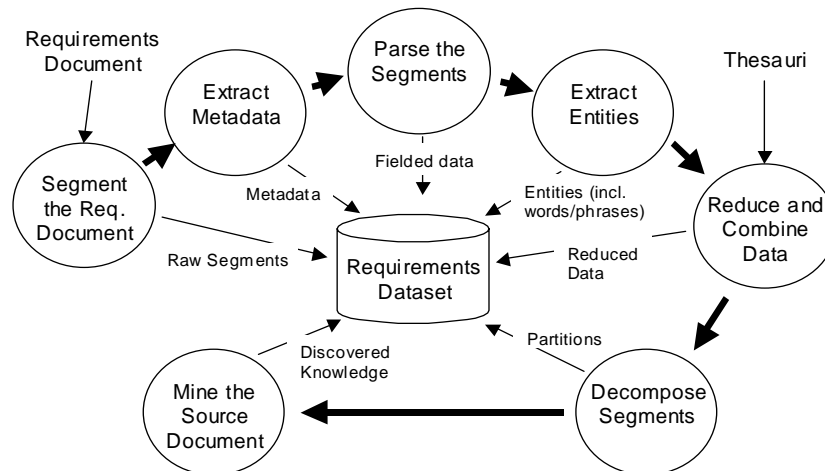


Figure 1. Process of mining knowledge from requirements documents

Segment the Requirements Document. If the source document is free text, a unit of analysis is needed to divide the source into meaningful smaller units or “segments” (e.g., sentences, paragraphs, sections, or small documents). By dividing the source into segments, we can derive meaning from the co-occurrence of phrases (or other data) in the segments. The size of the segment has a direct impact on the strength of association that one can assume exists between two phrases that co-occur in the segment – the smaller the segment, the stronger the presumed association.

Extract Metadata. Metadata is “data about data.” In this context, metadata about the segments can be extracted from the broader source document. For example, if the source has a hierarchical outline form, then metadata about the parent sections can be associated with the segment. So a requirements paragraph on “multi-spectral countermeasures” can have associated with it the fact that it occurs in a sub-section on “aviation” under the section “technology transition” in a document called “research plan.”

Parse the Segments. The segments themselves may have fielded data within them, such as allocated funding levels, points of contact, and supporting or dependent programs. These data can be parsed using delimiters or layout of the segment (e.g., “funding level” always follows the third blank line).

Extract Entities. This step of the process extracts words/phrases and other interesting data (“entities”) from free text fields.

Reduce and Combine Fields. The rich variety of expression in written language offers a challenge to the task of computer-based mining knowledge from text. Similar concepts can be expressed in many different ways. Effective analysis requires identifying terms that are similar in meaning in a given context and combining or associating those terms for subsequent analysis.

Decompose Segments. Within a given field, the co-occurrences of words/phrases are statistically analyzed to decompose or partition the segments into unique “buckets.”

Clusters of words/phrases that tend to occur together in the segments define these buckets.

Mine the Source Document. In addition to single-field decomposition of the segments, multi-field analysis techniques can be applied to discover hidden relationships in the source document.

Figure 2 illustrates the activity of mining knowledge from requirements documents and open literature. Beginning with one or more partitions decomposed from the source document, a query is developed and applied to open literature databases, and a dataset of open literature is created and mined to extract knowledge for the Program Manager. Each step is explained briefly in the following paragraphs.

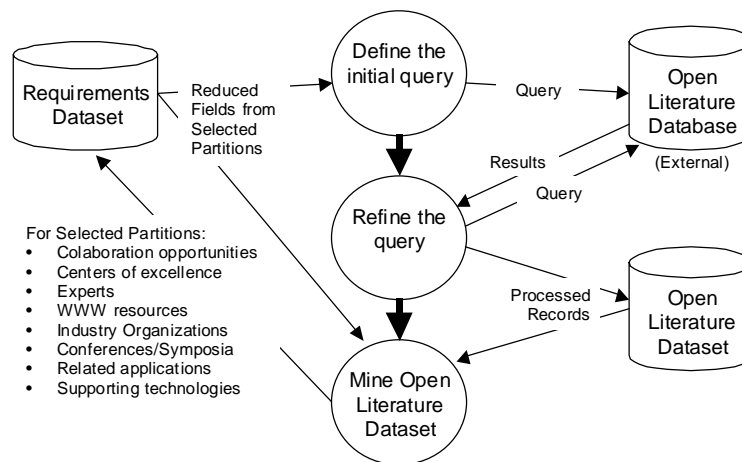


Figure 2. Mining knowledge from requirements documents and open literature

Define the initial query. The user selects areas of interest (“partitions” from the decomposition), and the tool helps compose an initial query for the open-source literature database (e.g., INSPEC, EI Compendex, U.S. Patents, Science Citation Index, etc.).

Refine the query. Using the results returned from initial query, the tool assists the user in an iterative sequence to refine the query. This helps ensure that the final open-source dataset is relevant to the areas of interest and that the ensuing knowledge-mining activity discovers applicable knowledge.

Mine the open literature dataset. As part of this process, a new dataset is created from the results of the query. The creation of this open-source dataset uses the same processes used to process the original requirements document (see Figure 1), except that the “segments” are “records” from the open-source literature database.

Results and Discussion

The following sections outline the results of our efforts toward each of the activities described above.

Segment the Requirements Document

The original TechOASIS import engine that was available at the beginning of this program was developed for well-structured flat text raw data files. It took advantage of some assumptions about the raw data, such as record- and field-delimiters occurring at line-breaks and item-delimiters being un-nested. This program pointed us toward more diverse and less-structured types of data. Therefore, the task of making the import engine more general and more powerful was among the first activities we approached.

As we looked into what would be necessary to accomplish this, we quickly realized that this would eventually be very closely associated with our approach to some of the other activities – specifically “Parsing the Segment”, “Extracting Metadata”, and “Extracting Entities.” We decided we should modify the overall architecture of our tools for creating and editing the import filters for raw data.

The end result is the Import Engine Editor.

Among the features added for “Segmenting the Requirements Document” are:

- Enable segmentation by paragraphs (Task C.1.1)
 - Conversion to the “Import Engine Editor” with a full Regular Expression (RegEx) parser enables byte-level delimiters (e.g., New Line and/or Carriage Return characters)
- Retain raw segment for access and display (Task C.1.2)
 - Made retention of the full raw record the default for all import filters
- Enable segmentation by document structure (Task C.1.3)
 - Also enabled by conversion to full RegEx parser. This allows dividing a document according to the “style” of the header text (e.g., use of all caps at certain outline levels, or numbered outline headings such as 1.2.1, and 2.a.4.i) by creating a pattern that matches the various header styles. [See SPS Document section titled “ENABLE SEGMENTATION BY DOCUMENT STRUCTURE”]
- Character-based record and field delimiters import, enabling import of SGML and XML tagged data (Task C.1.4)
 - Also enabled by conversion to full RegEx parser that allows matching of the XML and SGML tags anywhere in the raw data file.

Note: Please refer to the following document(s) for full descriptions of these capabilities:

- VantagePoint User’s Guide (IMPORT ENGINE, IMPORT ENGINE EDITOR)
- VantagePoint On-line help (IMPORT ENGINE, IMPORT ENGINE EDITOR)
- SPS Document (IMPORT ENGINE, IMPORT ENGINE EDITOR, ENABLE SEGMENTATION BY DOCUMENT STRUCTURE)

Segmentation by user selection: Another approach to segmenting raw data files (e.g., the requirements document) is to provide a capability for the user to interactively affirm or correct the Import Engine's selection of the segment start/end points.

- Enable segmentation by user selection (Task C.1.1)
 - A “flag” was added to the Import Engine Editor and an interactive dialog was created in the Import Engine for this purpose.
 - Third-party software tools for segmenting PDF documents. After extensive investigation and evaluation, we proposed that third-party tools would be the most cost effective approach to converting PDF documents to plain text and for user selection of the segments. The following tools were recommended and selected. [See the SPS Document section titled “PDF-TO-TEXT” for a full discussion of these tools and their use in segmentation by user selection.]
 - Adobe Acrobat v5.0
 - iStructure
 - BCL Jade (software plug-in for Adobe Acrobat)

Note: Please refer to the following document(s) for full descriptions of these capabilities:

- SPS Document (PARSE SEGMENTS – USER CONFIRMATION SCREENS)
- SPS Document (PDF-TO-TEXT)

Segmentation by content measures: The task of parsing segments from a document might also use “shift in discourse” to determine segment boundaries. This capability was named “enable segmentation by content measures” for this project (Task C.1.3). For this activity, the following functions were created:

- Similarity of adjacent segments – Record Similarity Dialog (for researching content measures)
- 3 correlation measures for similarity of adjacent records – Pearson, Cosine, and Max Proportional
- Byte/word counts for each segment in the preview panes

Note: Please refer to the following document(s) for full descriptions of these capabilities:

- SPS Document (DOCUMENT SEGMENTATION BY CONTENT MEASURES)

PDF Import: The project plan anticipated that some of the requirements documents might be in Portable Document Format or PDF (Task C.1.4). The third-party tools listed earlier were recommended and selected for this purpose. The use of these third-party tools is described in their respective documentation.

Note: Please refer to the following document(s) for full descriptions of these capabilities:

- SPS Document (PDF-TO-TEXT)

Extract Metadata

The activity of developing the capability to extract metadata from requirements documents (Task C.2) resulted in three functions:

- Import Variables – These provide “variables” to hold values (i.e., text strings) and locations (i.e., where the text string occurs in the requirements document) of specific items extracted during import. Import Variables are used when metadata for segments occur outside the segment boundaries. For example, the section and chapter headings (the metadata) for paragraphs (the segments) usually occur outside of the paragraphs. Import Variables provide a mechanism for capturing the section and chapter headings and associating them with each of the paragraphs in the section or chapter.

Note: Please refer to the following document(s) for full descriptions of Import Variables:

- VantagePoint User’s Guide (IMPORT ENGINE, IMPORT ENGINE EDITOR)
 - VantagePoint On-line help (IMPORT ENGINE, IMPORT ENGINE EDITOR)
 - SPS Document (PASS METADATA INTO RECORDS (IMPORT VARIABLES))
- FrankenRecords – Other databases may also be a source of metadata. For example, consider a database of Science and Technology Objectives (STOs). The primary database might include an abstract of the Objective, information about the primary contacts at the responsible organizations, the Platforms supported by the Objective, etc. Another database might contain the annual funding profiles of various Program Elements that support each STO – a many-to-one relationship. These funding profiles can be considered metadata about the STO. Frankenrecords were developed to “stitch together” records from diverse data sources based on a common field (in the example, the STO number).

Note: Please refer to the following document(s) for full descriptions of FrankenRecords:

- VantagePoint User’s Guide (FRANKENRECORD)
 - VantagePoint On-line help (FRANKENRECORD)
 - SPS Document (PASS METADATA INTO RECORDS (COMBINE RECORDS))

Parse the Segments

Parsing fielded data from the segments is the next step in the process (Task C.3). Some of the capabilities necessary for this step existed in TechOASIS prior to the beginning of

this Phase II effort. The effort of re-building the Import Engine and creating the Import Engine Editor were also aimed at this activity.

- Import Engine and Import Engine Editor – The Field Definition portion of the Import Engine Editor contains the tools to parse many types of data from the segments. The “toolbox” contains 15 different commands with a rich set of parameters, and each Field Definition contains a sequence of commands (a “Command Stack”) that parses the data from the segment.

Note: Please refer to the following document(s) for full descriptions of Field Definition commands in the Import Engine and Import Engine Editor:

- VantagePoint User’s Guide (IMPORT ENGINE EDITOR)
 - VantagePoint On-line help (IMPORT ENGINE EDITOR)
- Parse segments Optimal Approach – During the course of this program, we have investigated and evaluated a variety of approaches to parse fielded data from the segments. The one selected for this activity uses an iterative approach. The essence of the approach is for the user to (1) select a set of terms in a field of interest (i.e., make a “group” of “terms of interest”) and (2) apply the “Extract Nearby Phrases ...” tool to a free text field using the group of terms as a “seed”. “Extract Nearby Phrases ...” will run the NLP parser on sentences surrounding any occurrence of any of the terms and create a field of the noun phrases in those sentences. This provides a contextually rich set of phrases closely related to the terms of interest.

Note: Please refer to the following document(s) for full descriptions of “Extract Nearby Phrases”:

- VantagePoint User’s Guide (EXTRACT NEARBY PHRASES)
- VantagePoint On-line help (EXTRACT NEARBY PHRASES)
- SPS Document (PHRASES IN THE CONTEXT)

Extract Entities

This project added several capabilities to identify and extract entities from segments of text.

- Entity extraction using dictionaries (Task C.4.2) – The Field Definition portion of the Import Engine Editor contains a command for Entity Extraction. The entity dictionary (a text file) contains a list of terms or Regular Expressions that the Import Engine should save in the field if found in the text.

Note: Please refer to the following document(s) for full descriptions of Entity Extraction commands in the Import Engine and Import Engine Editor:

- VantagePoint User’s Guide (IMPORT ENGINE EDITOR)
- VantagePoint On-line help (IMPORT ENGINE EDITOR)
- SPS Document (ENTITY EXTRACTION)

- Entity Extraction User Confirmation Screens (Task C.4.2) – As with record start/end, it is occasionally difficult to specify data elements perfectly in the import filter. User confirmation screens for extracting data elements have been built to help in difficult situations. This confirmation setting is made in the Import Engine Editor under the “Field Settings” tab in “Confirm Entities on Import”.

Note: Please refer to the following document(s) for full descriptions of Entity Extraction User Confirmation Screens commands in the Import Engine and Import Engine Editor:

- VantagePoint User’s Guide (IMPORT ENGINE EDITOR)
 - VantagePoint On-line help (IMPORT ENGINE EDITOR)
 - SPS Document (ENTITY EXTRACTION USER CONFIRMATION SCREENS)
- Protect extracted entities during NLP parsing (Task C.4.2) – Entity extraction dictionaries may be applied during the Import Engine’s NLP phrase-parsing step. In this case, the entities are extracted and the NLP parsing runs on the remaining text. The result is a field containing the entities and the phrases that remain after the entities are extracted.

Note: Please refer to the following document(s) for full descriptions of commands for “Entity extraction with NLP parsing” in the Import Engine and Import Engine Editor:

- VantagePoint User’s Guide (IMPORT ENGINE EDITOR)
 - VantagePoint On-line help (IMPORT ENGINE EDITOR)
 - SPS Document (ENTITY EXTRACTION WITH NLP PARSING)
- Entity Dictionaries are usually specific to a particular purpose. However, two general entity dictionaries have been developed and deployed during this project:
 1. Countries – derived from several sources and augmented to include possessives; and
 2. Dates – which match several common date formats.
- Find Entities consisting of general sequences of terms (e.g., “Generalized Episodes” – Task C.4.2) – The final implementation of this capability is in an enhanced “Find” dialog that can be used when viewing a list of a free text field or the raw record field. The “Find” dialog contains three stages of the query, four Boolean and ten proximity operators, and whole word or sub-string matching (manages sub-string matches such as “sea” in “research”). Combining the 14 operators with the richness of Regular Expression matching provides a mechanism for finding “episodes” – loosely structured “chunks” of text – in free text. The Boolean operators provide for co-occurrence style searching. The order-dependent proximity operators (Followed by, Followed

by Adjacent, Followed by Near2, etc.) enable searching for sequences. The proximity operators (Near2, etc.) help in situations where the order may be less predictable.

Note: Please refer to the following document(s) for full descriptions of the “Find Dialog”:

- SPS Document (ENHANCED FIND DIALOG FOR GENERALIZED SEQUENCES AND PATTERNS)

Reduce and Combine Data

Two tools form the core capability in TechOASIS for reducing and combining (or normalizing) data – (1) Thesaurus and (2) List Cleanup. The emphasis in this project has been extending and enhancing these tools to improve performance and provide greater flexibility. The following lists outline the improvements (Task C.5):

- List Cleanup
 - Find/Select All – to help with collecting similar items.
 - Sort/find/select/cut/paste
 - String and RegEx match/management in “Find Close Match” section of Confirmation dialog
 - Save/Resume List Cleanup operations
 - Use stemming in Cleanup
 - Use partial match thesaurus in Cleanup
- Thesaurus Editor
 - Full Interactive Cleanup Confirmation for Top-Level Items – provides a way to normalize a thesaurus that has been developed using List Cleanup.
 - Improved Performance – for non-RegEx and anchored RegEx entries
 - Sort/find/select/cut/paste
 - Identification and removal of redundant sub-items
 - Interactive resolution of identical sub-items assigned to more than one top-level item.
 - Case sensitivity
 - User-managed default choice for root items
- Thesauri software applications (Task C.5.2) – Adapted and delivered Roget Thesaurus.

Note: Please refer to the following document(s) for full descriptions of List Cleanup and Thesaurus Editor and the extensions added by this project:

- VantagePoint User’s Guide
- VantagePoint On-line help
- SPS Document (LIST CLEANUP AND THESAURUS ENHANCEMENTS)

Decompose Segments

The user's activity of decomposing (analyzing) the segments is aided by calculating and providing multiple diverse perspectives on the data contained in the segments. Three significant tools were developed to support this activity.

- Multiple Graphical Depictions of Field Data in Maps – Graphical pop-up displays combine information from one field (e.g., categorical data) on a map of data from a different field, providing the ability to compare and contrast the segments that fall into different clusters across several distinct fields. There are several depictions available (e.g., Pie Charts, Bar Graphs and 3D Bar Graphs, Box Plots, and Text), and default depictions can be assigned to specific data types.

Note: Please refer to the following document(s) for full descriptions of these mapping extensions added by this project:

- VantagePoint User's Guide (Maps and Map Preferences)
 - VantagePoint On-line help (Maps and Map Preferences)
 - SPS Document (POP-UP OF CATEGORICAL DATA ON MAPS)
- Statistical analysis of fields – An extension to the Detail Windows (described later) was developed to provide a statistical analysis of the fielded data. "Expectancy arrows" indicate how much the item's number of records in the detail window departs from expectation based on the item's number of records in the whole dataset.

Note: Please refer to the following document(s) for full descriptions of these Expectancy Arrows:

- VantagePoint User's Guide (Detail Windows)
 - VantagePoint On-line help (Detail Windows)
 - SPS Document (DETAIL WINDOW – PROBABILITIES)
- Additional domain-specific thesauri – The following thesauri and entity dictionaries were developed and deployed as a part of this project: (1) DoD Acronym thesaurus, (2) FCS Acronym thesaurus, and (3) entities derived from DoD/FCS dictionary and acronym lists.

Mine the Source Document

The activity of mining the source document uses analytical techniques to extract knowledge from the segments, which collectively make up the source document. TechOASIS contains several tools that support this activity. The following capabilities were added in the course of this project.

- Multi-field text analysis tools (Task C.7) – Some of the most powerful tools in TechOASIS are the mapping and decomposition tools – Factor maps (for clustering terms within a field), auto-correlation and cross-correlation maps (for visualizing the relationships among individual items within a field), and

the Principal Components Decomposition. With the exception of cross-correlation maps, these are single-field analytical tools (cross-correlation maps use a second field to determine the relationships). The multi-field text analysis capability added by this project enables the user to create each of these analytical products (maps and decompositions) using clusters or groupings developed in another field, using a subset of the full dataset that may be the result of an analysis in another field.

Note: Please refer to the following document(s) for full descriptions of this capability:

- SPS Document (CREATING MAPS USING SUBSETS OF RECORDS)
- Clustering techniques (Task C.7) – Building on a recommendation from our Phase I results, a new clustering technique was added for evaluation and experimental purposes. The CLUTO clustering toolkit has been developed by Professor George Karypis at University of Minnesota, Department of Computer Science. The usage terms of CLUTO allow it to be freely used for research purposes by U.S. Government agencies, and Dr. Karypis has granted permission to use CLUTO in this project. We have incorporated the CLUTO link-clustering toolkit into Tech OASIS as a foundation to explore the capabilities of this data mining technique for our data. The interaction with the CLUTO link-clustering tool suite is currently via a VB-script command (Dataset.ClusterRecords).

Note: Please refer to the following document(s) for full descriptions of this capability:

- SPS Document (CLUTO, A CLUSTERING TOOLKIT)
- “CLUTO – A Clustering Toolkit.” University of Minnesota.
- Three-Field Visualization (Task C.7.1) – Two capabilities were added to address this requirement. (1) Detail Windows display addition field(s) in “dockable” windowpanes adjacent to the main analytical view. When viewing a co-occurrence matrix, this constitutes a three-field view. (2) A macro- and menu-driven command was added to export data from a co-occurrence view in a format that can be used in Microsoft Excel to create Pivot Charts and Pivot Tables.

Note: Please refer to the following document(s) for full descriptions of these capabilities:

- SPS Document (DETAIL WINDOWS)
- SPS Document (EXPORT 3-FIELD CO-OCCURRENCE MATRIX)

Refine the Query

The original intent of the tasking (C.8) for this activity was to improve the trans-lingual functionality of TechOASIS Query Refinement. Between project initiation and beginning work on this task, the translation capability of TechOASIS became inoperative

due to an extensive change by SYSTRANSOFT in the API for their product. No project resources were available to rebuild the interface between SYSTRAN's product and TechOASIS. Therefore we refocused this activity to improve Query Refinement for English-language datasets.

Note: Please refer to the following document(s) for full instructions on how to use Query Refinement in TechOASIS:

- TOAS Final Technical Report, Jul 2001

The improvement added in this project provides a tool for improving the selection of the candidate query terms using a variation of the Term Frequency, Inverse Document Frequency (TFIDF) metric commonly used for information retrieval (IR). Many references describing TFIDF are readily available (for example, see Jing, et al., 2002). We use TFIDF as a metric to quantitatively rank terms in a dataset based on their capacity to differentiate records based on their presence or absence in the records. For our purposes, Term Frequency (TF) is the number of times the term appears in the dataset, also referred to as "Number of Instances" in TechOASIS. (In IR, it means the number of times the term appears in a given document.) Inverse Document Frequency is the log of the ratio of the total number of records and the number of records in which the term appears. If a term appears in all records, IDF is zero ($\log(1)$).

For Query Refinement, the TFIDF metric can be used to narrow and focus the term-space that is used when rating records. Terms with higher TFIDF values provide better discrimination among the records, and a narrower term-space will improve performance and should provide faster convergence to suggested refinements.

A four-step process is used to select terms using TFIDF.

1. A group is created in a "unique record pointer" field, such as the Raw Record, Accession Number, or key field. This group should contain all of the records.
2. A TFIDF matrix is created using the "Term Field" containing the terms of interest as rows, and the grouped records (from the "unique record pointer" field). See the User's Guide or On-Line Help for detailed instructions.
3. Sort the single-column matrix and select the top terms for the term-space. There currently is no agreed-to rule of thumb for where to cut off the term-space. Several approaches have been tried; among them (a) the square root of the maximum value, (b) $< 50\%$ of the terms with the cutoff at a breakpoint, and (c) plotting the values and visually determining a break point. This is clearly an area deserving further research.
4. Create a group in the "Term Field" using the selection in the matrix (right-click menu, "Add Row Selections to Group").

This group is then selected as the "term-space" in the second stage of Query Refinement.

Note: Please refer to the following document(s) for full descriptions of these capabilities:

- VantagePoint User's Guide (on how to create a TFIDF matrix)
- VantagePoint On-line help (on how to create a TFIDF matrix)
- TOAS Final Technical Report, Jul 2001 (on how to use Query Refinement in TechOASIS)

Mine Open Literature Dataset

The final step in the process is to use the results of the analysis of requirements documents to mine datasets drawn from the open literature to find research results and other information that might be drawn in to help Army research centers meet the operational requirements.

For a period of time during the final 18 months of the project, the target 'use-environment' for this program became mining/classifying Science/Technology Objectives (STOs) based on an analysis of Operational Requirements Documents (ORDs – the 'requirements' documents in our discussion). In this environment, the STOs became the parallel to the open literature – the relationship being that instead of mining published R&D results, we would mine descriptions of on-going Army research projects to find activities that might support multiple (and perhaps previously unassociated) requirements and find requirements that currently have no research support.

The aim became improving the classification performance of existing algorithms via term expansion/combination. At a very practical level, the difficulty quickly became that there are differences in the types of language used in STOs and ORDs – requirements are frequently expressed using different kinds of words than are used to describe research. This pointed to thesauri techniques, and our specific approach sought to apply a network thesaurus (e.g., Princeton's WordNet) to try to discover the associations.

The WordNet data files contain over 115,000 "synsets" – sets of synonyms with a common meaning/usage. For example, there are 18 synsets in the "noun" and "verb" databases that include the word "force." The following two tables summarize these synsets.

9 senses of "force" (noun)

Sense 1: military unit, military force, military group, force -- (a unit that is part of some military service; "he sent Caesar a force of six thousand men")

Sense 2: power, force -- (one possessing or exercising power or influence or authority; "the mysterious presence of an evil power"; "may the force be with you"; "the forces of evil")

Sense 3: force -- ((physics) the influence that produces a change in a physical quantity; "force equals mass times acceleration")

Sense 4: force, personnel -- (group of people willing to obey orders; "a public force is necessary to give security to the rights of citizens")

Sense 5: force -- (a powerful effect or influence; "the force of his eloquence easily persuaded them")

Sense 6: violence, force -- (an act of aggression (as one against a person who resists); "he may accomplish by craft in the long run what he cannot do by force and violence in the short one")

Sense 7: force, forcefulness, strength -- (physical energy or intensity; "he hit with all the force he could muster"; "it was destroyed by the strength of the gale"; "a government has not the vitality and forcefulness of a living man")

Sense 8: force -- (a group of people having the power of effective action; "he joined forces with a band of adventurers")

Sense 9: effect, force -- ((of a law) having legal validity; "the law is still in effect")

9 senses of "force" (verb)

Sense 1: coerce, hale, squeeze, pressure, force -- (to cause to do through pressure or necessity, by physical, moral or intellectual means : "She forced him to take a job in the city"; "He squeezed her for information")

Sense 2: impel, force -- (urge or force (a person) to an action; constrain or motivate)

Sense 3: push, force -- (move with force, "He pushed the table into a corner")

Sense 4: force, thrust -- (impose or thrust urgently, importunately, or inexorably; "She forced her diet fads on him")

Sense 5: wedge, squeeze, force -- (squeeze like a wedge into a tight space; "I squeezed myself into the corner")

Sense 6: force, drive, ram -- (force into or from an action or state, either physically or metaphorically; "She rammed her mind into focus"; "He drives me mad")

Sense 7: force -- (do forcibly; exert force; "Don't force it!")

Sense 8: pull, draw, force -- (cause to move along the ground by pulling; "draw a wagon"; "pull a sled")

Sense 9: storm, force -- (take by force; "Storm the fort")

These are networked relationships – many-to-many. A word frequently has multiple meanings (synsets) and a synset frequently has more than one word.

Using the WordNet resource files, a TechOASIS thesaurus was developed (filename *WordNetWordBoundaryFrontWordBoundaryBackLE5.the*). For illustration, the following is a small excerpt showing the 14 multi-term synsets that include the word 'force' (single term synsets are omitted here). The thesaurus 'aliases' (e.g., '01484954 35 v') are WordNet's coding structure for synsets.

**00722891 32 v	**01543295 35 v	**04766490 07 n
100 1 \bforce\b	100 1 \bforce\b	100 1 \bforce\b
100 1 \bthrust	100 1 \bstorm\b	100 1 \bforcefulness
**00907783 04 n	**01603542 36 v	100 1 \bstrength
100 1 \bforce\b	100 1 \bforce\b	**07701234 14 n
100 1 \bviolence	100 1 \bimpel\b	100 1 \bforce\b
**01406785 35 v	**01817891 38 v	100 1 \bmilitary force
100 1 \bdraw\b	100 1 \bforce\b	100 1 \bmilitary group
100 1 \bforce\b	100 1 \bpush\b	100 1 \bmilitary unit
100 1 \bpull\b	**02429697 41 v	**07710741 14 n
**01474193 35 v	100 1 \bcoerce	100 1 \bforce\b
100 1 \bdrive\b	100 1 \bforce\b	100 1 \bpersonnel
100 1 \bforce\b	100 1 \bhale\b	**09780795 18 n
100 1 \bram\b	100 1 \bpressure	100 1 \bforce\b
**01484954 35 v	100 1 \bsqueeze	100 1 \bpower\b
100 1 \bforce\b	**04586702 07 n	
100 1 \bsqueeze	100 1 \beffect	
100 1 \bwedge\b	100 1 \bforce\b	

The resulting thesaurus is massive:

- Size 6.3 MB
- Number of lines: ~ 318,000
- Number of root items: ~115,000

Several compromises were necessary to achieve acceptable levels of performance. (In this context “performance” means “matching appropriate text strings.”) Regular Expression commands are added to improve the precision of the matches. The leading ‘\b’ is common to the entire thesaurus, forcing all matches to occur on a leading word boundary. Note that this compromise improves the performance somewhat, but at the expense of missing terms with prefixes. A trailing ‘\b’ is used for any term less than six characters in length (empirically determined) to minimize imprecise sub-string matches. Note that this improvement in the matching comes at the expense of missing some terms with suffixes.

Another component of “performance” is speed. Applying this thesaurus to a list of ~5,000 phrases takes several days on an average desktop computer. The performance can be improved significantly (to a few hours) if the thesaurus is changed to anchor the sub-items at the front (RegEx ‘^’) or the end (RegEx ‘\$’); however, this improvement comes with the price of missing word matches embedded in multi-word phrases or single words with prefixes or suffixes.

The basic process for using this thesaurus follows:

1. Import NLP Phrases, but parse the multi-word phrases into individual words. This eliminates the need for sub-string matching and permits anchoring the thesaurus sub-items at the front (however, still missing prefixes).
2. Clean the words, typically accepting the default clean-up, although the user is free to verify the clean-up step.
3. Apply a standard stop-words thesaurus. This mainly cleans up numbers and uninteresting words (again, can be tailored by the user).
4. ‘Create Groups Using Thesaurus’ using the WordNet thesaurus (filename *WordNetWordBoundaryFrontWordBoundaryBackLE5.the*). We use ‘Create Groups Using Thesaurus’ rather than regular ‘Thesaurus’ to allow words in the list to be grouped to more than the one synset encountered in the thesaurus.
5. Run a macro that finds and combines equivalent groups (filename *CombineSubSetGroups-NameWithItemsA.vpm*).
6. Create a Thesaurus using these groups.
7. Apply the thesaurus to the original list to create the reduced list for analysis.

This sequence is illustrated in the following paragraphs.

We began with a dataset of 38 records selected from a larger sample set, and performed a standard cleanup on the noun phrases. The following table shows a small portion of the resulting field.

# Records	# Instances	Abstract (NLP) (Phrases) (Cleaned)
10	16	haptic interface
9	10	user
7	9	system
6	6	development
6	7	haptic device
6	10	operator
5	6	device
5	5	effectiveness
5	6	force
5	7	method
5	6	motion
5	5	one
4	4	applications
4	4	combination
4	4	dynamics
4	4	experimental results
4	5	friction
4	4	haptic rendering
4	4	presented
4	4	virtual environment
4	5	virtual objects
		(... 885 items in all)

We then parse the phrases into words. This has the effect of reducing and aligning the ‘feature’ space for the analysis. This will enable some degree of relationship between, for example, (a) a record that contains ‘haptic interface’ and not ‘haptic device’ and (b) a record that contains ‘haptic device’ and not ‘haptic interface.’ Furthermore (and more to the point for the current discussion), a single-word term is much more likely than a multi-word phrase to have a match in the WordNet thesaurus. The following table shows the “top” (# Records > 10) portion of the resulting field.

# Records	# Instances	Abstract (NLP) (Phrases) (2) (Cleaned)
33	95	Haptic
23	38	Device
22	58	Force
22	44	System
17	31	Interface
16	29	Model
15	38	Objects
14	22	User
14	27	Virtual
12	14	Two
11	25	Interaction
10	23	Based
10	10	Development
10	20	Display
10	12	Environment
10	20	Method
10	16	Motion
10	11	New
10	18	Operator
		(... 752 items in all)

This list has 19 terms appearing in 10 or more records.

Applying the WordNet thesaurus to create groups, our list of 752 items matched 4,114 ‘synsets’ creating 4,114 groups (sample shown in the following illustration).

	# Records	# Instances	Abstract (NLP) (P)	06236902 10 n	00722891 32 v	00907783 04 n	01310566 35 v	01406785 35 v	01474183 35 v	01484954 35 v	01543295 35 v	01603542 36 v	01817891 38 v	02429697 41 v	04766400 07 n	04899893 07 n	07701234 14 n	07710741 14 n	09780795 18 n	10719108 19 n	00023103 03 n
1	22	58	force																		
2	3	3	strength																		
3	1	1	power																		
4	1	1	pull																		
5	1	1	push																		
6	33	95	haptic																		
7	23	38	device																		

Note the explosion (> 5x in this example) of the number of items resulting from the many-to-many mapping of terms-to-synsets in the WordNet thesaurus.

Running a script to combine 100% subsets, this reduces to 571 term combinations (sample shown in the following illustration).

	# Records	# Instances	Abstract (NLP) (Phrase)	following	Food	footprint	force pull	force push	force strength	force power	shape work form	spring form
1	22	58	force									
2	3	3	strength									
3	1	1	power									
4	1	1	pull									
5	1	1	push									
6	33	95	haptic									

Creating a field from the group names results in a field of 571 items (sample shown in the following illustration).

	# Records	# Instances	Abstract (NLP)
1	26	72	force effectiveness effector
2	23	59	force pull
3	23	59	force push
4	23	61	force strength
5	22	59	force power
6	34	99	haptic tactile HapticFlow
7	23	38	device
8	23	50	system scheme
9	21	45	model simulation
10	18	32	interface port
11	17	31	model framework
12	16	31	model pattern
13	16	31	model posture
14	15	17	new novel
15	15	38	objects

The following table shows the “top” (# Records > 10) portion of the resulting field.

# Records	# Instances	Abstract (NLP) (Phrases) (2) (Cleaned) (Group Names)
34	99	haptic tactile HapticFlow
26	72	force effectiveness effector
23	38	Device
23	59	force pull
23	59	force push
23	61	force strength
23	50	system scheme
22	59	force power
21	45	model simulation
18	32	interface port
17	31	model framework
16	31	model pattern
16	31	model posture
15	17	new novel
15	38	Objects
14	18	one 1 10 100 1D single
14	24	results effectiveness effector
14	20	two 2 25 2D
14	22	User
14	28	virtual practical
13	28	contact touch
13	21	motion movement
13	16	results resolution solution
11	13	environment surroundings
11	20	information data
		(... 571 items in all)

This list has 38 items appearing in 10 or more records, compared to 19 in the original cleaned list – the end result being that using the WordNet thesaurus in this way increases the term frequency and reduces the number of terms as originally intended. However it does so at the expense of specificity of the narrower terms. The term ‘strength’ (3 records) gets combined with the term ‘force’ (22 records), resulting in the term ‘force | strength’ (23 records). Similarly the term ‘power’ (1 record) gets combined with the term ‘force,’ resulting in the term ‘force | power’ (22 records).

The conclusions from this activity are mixed.

- (+) Using a network thesaurus such as WordNet does appear to hold promise for term association between datasets with differences in the kind of words that are used (e.g., requirements vs. research objectives).
- (-) There are several compromises that must be made to achieve even minimally acceptable levels of performance (both matching and speed) on datasets of modest size.

The end products of this activity provide a good basis for continued research on this problem.

Automation/Macros

The Visual Basic Scripting capability of TechOASIS has proven to be one of its strongest points. With scripting a user is able to customize TechOASIS to their analytical approach and speed up the analytical process.

Part of this project was to add scripting commands for as many of the added capabilities as practical. Scripting does not benefit some capabilities, because they require too much user interaction and judgment. However, many functions can benefit, and the following were added during the course of this project:

- App
 - GetDetailWindowNames
 - ImportFile
 - SelectDetailWindow
 - ShowDetailWindow
- Dataset
 - ClusterRecords
 - CreateKeyField
 - GetDatabaseNames
 - GetNumTitleViewRows
 - GetNumTitleViewRowsSelected
 - GetActiveViewName
 - GetDataType
- List
 - FindSelectRegex
 - FindSelectRegexPassive
 - GetCurrentCell
 - GetSelectedRows
 - GetSelectedCols
 - GetRanges
 - GetNumberOfMembersInGroup
 - GetNumberOfRecordsInGroup
- View
 - CreateDetailWindow
 - CreateMatrixInstances
 - CreateAutocorrelationMapFromRecordSubset
 - CreateCrosscorrelationMapFromRecordSubset
 - CreateFactorsMapFromRecordSubset
 - CreateCrosscorrelationMapInst
 - CreateAutoCorrelationMatrix
 - CreateCrossCorrelationMatrix
 - CreateIndirectLinksMatrix
 - CreateTFIDFMatrix
- Matrix
 - Export3Dmatrix
 - GetCurrentCell
 - GetSelectedRows
 - GetSelectedCols
 - GetRanges
- Detail
 - Copy
 - ZoomOut
 - ZoomOutAll
 - IsList
 - IsChart
 - Expand
 - Contract
 - DeselectAll

- SelectAll
- ShowChart
- ShowList
- SortByNumRecs
- SortByProbability
- SortByItemLabels
- SelectField
- SaveAsJPEG
- SaveAsBMP
- SetChartStyle
- MetaTagSetChartData
- GetNumRows
- SelectRows
- FindSelect
- GetValue

Note: Please refer to the following document(s) for full descriptions of these capabilities:

- VantagePoint On-line help (Automation and Scripts / VantagePoint Scriptwriter Reference)

***Note:** The preceding sections describe the results of the project as originally conceived and funded. During the course of Phase II, two contract modifications (Phase II Plus) added targeted activities to the work scope to address issues and capabilities that benefited the Army as well as other co-funding agencies. The following sections describe the results of these added activities.*

Leverage Open-Source Data Repositories

In parallel with the activity to investigate mining open literature databases (i.e., fee-based resources), tasks were added via Phase II Plus to research automating the computer interface with and the download from freely available, open web sites of research literature. Government agencies provide these types of services (e.g., DOE/OSTI and DTIC/STINET), as do some professional organizations (e.g., IEEE) and ongoing commercial research projects (e.g., at the time NEC's CiteSeer or ResearchIndex). The goal of these tasks was (a) to evaluate the effectiveness of these no-cost resources compared with the fee-based services and (b) to recommend an appropriate role for each type of data source.

Somewhat unanticipated was the degree of resistance the host organizations would present and their unwillingness to collaborate on this research. In general, their concerns were that opening the door to large-scale retrieval of data might tax their systems and deny service to their broader constituency. Some sites (e.g., IEEE and the U.S. Patent Office) administratively restrict the use of "crawlers" on their sites through their "Terms of Use" policies. While some of these concerns are legitimate, we sought only to research the potential of the concept. Much of the tasks' resources were expended researching the terms-of-use, contacting and negotiating with five organizations, and

eventually reaching essentially a dead end with each. In some cases, we were unable to engage sufficient interest to gain the necessary support. In others, especially the government sites, the final offer of collaboration was too limited to be of much value.

With the sponsor's support, we proceeded with very limited retrieval from DTIC/STINET site using a crawler based at Sandia National Laboratory as proof-of-concept. The import filter for this data source (STINET.conf) is included with the software distribution. This demonstrated that TechOASIS could work with crawlers and import the resulting data. However, the volume of the retrieved data was insufficient to perform an evaluation of the effectiveness of the data source.

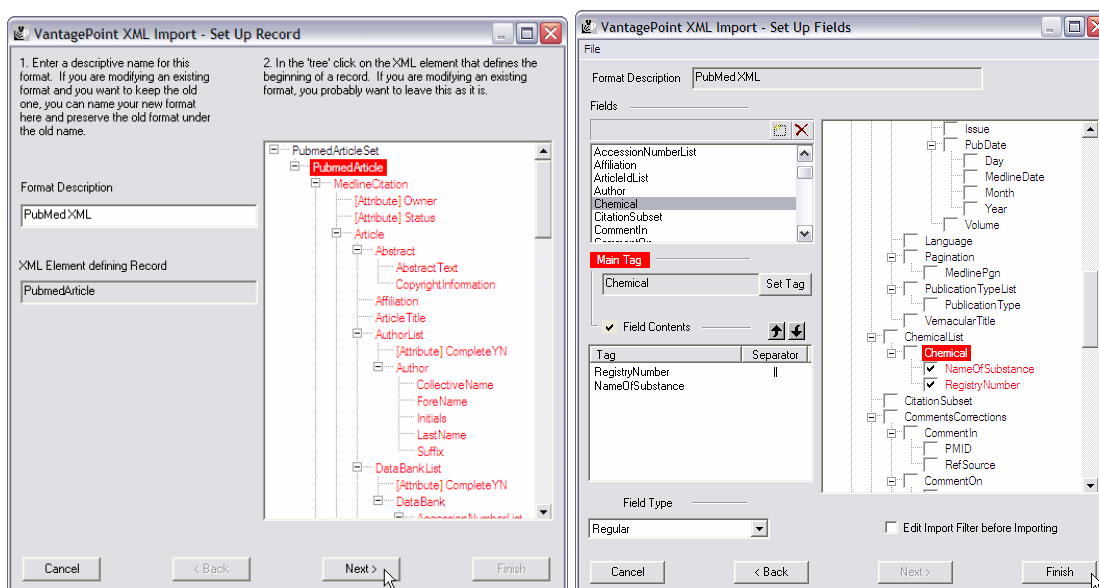
Later, in conjunction with another Phase II Plus task (see "WebQL interface" below), we were able to continue this activity. Using data provided by the Army through their WebQL license and in collaboration with our sponsor's technical representative, a study compared two data sources – IEEE Conferences (available via the Internet) and EI Compendex (a fee-based service) – and produced a recommended practice for using fee-based-services to profile free-text information (see Appendices A and B).

Quick XML Import

Earlier developments in this project enabled importing of XML data (see earlier sections on changes and enhancements to the Import Engine for character-based record delimiters, field tags, and field delimiters). The expanding use of XML for text databases in the TechOASIS user community and the rigorous data structures that typify XML data sources create an opportunity to streamline the import of XML data.

The original conceptual design for "Quick XML Import" planned to use the XML DTD or Schema as the primary source of information about the structure of the XML data. Several alternatives were evaluated and an initial prototype was trialed. As alternative XML data sources were explored, it became clear that many XML data sources either do not use/maintain the DTD/Schema, or do not routinely distribute them. Furthermore, many sources of XML data are not "well formed" – which means it "does not strictly adhere to the standards in the XML standards community."

To provide broader coverage of potential sources, we adopted an approach that examines the structure of the actual XML data (rather than the DTD or Schema) to determine the data structure. This approach requires the user to go through a two-step process to indicate the record and field tags the first time each type of XML data is imported. The format is then saved in a small resource file and may then be shared with others. Optionally, the user may use the Import Engine and Import Engine Editor for subsequent processing after the XML data is imported.



Note: Please refer to the following document(s) for full descriptions of these capabilities:

- SPS Document (QUICK XML IMPORT)

Simultaneous Cross Dataset Analysis

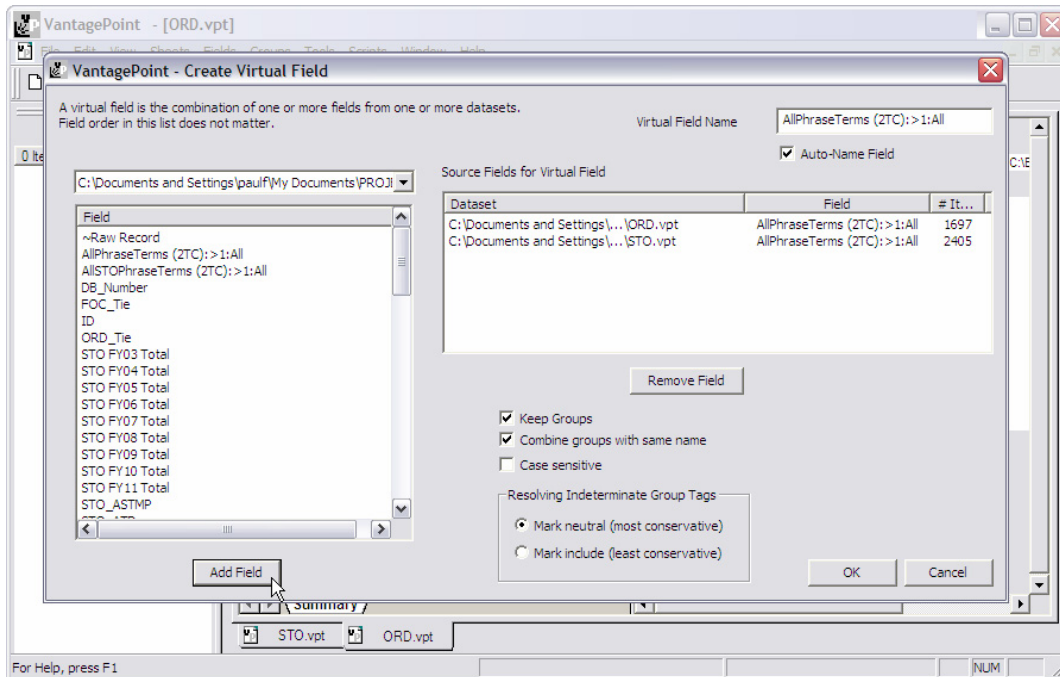
During the course of Phase II, feedback from users pointed toward a requirement to collectively analyze diverse datasets. For example, one group of users wanted to find clusters of similar records from four distinct data sources: (1) a relatively straightforward requirements dataset, (2) a dataset of performance parameters, (3) a dataset of operating capabilities, and (4) a more complex research objectives dataset that included dozens of fields such as funding information by year.

This problem is different than other problems addressed earlier in this project. The earlier work addressed using the results of an analysis of one dataset to direct the analysis of a different dataset. This later requirement seeks to analyze the datasets together as a group.

An initial approach fused the datasets into one very large dataset. This was successful; however, the process of fusing the datasets was laborious and created a massive data file that was inefficient to work with because of the resulting size of the collective data structures. The users advocated for a new approach, and a contract modification funded this activity.

This capability is aimed at conducting a collective analysis of two or more datasets without requiring the user to fuse the datasets together. The implementation enables the user to form 'virtual fields' in a new type of dataset. For a given analysis the user will create a virtual field for each of the fields necessary for the analysis. For a single field analysis (e.g., lists, auto-correlation matrix or map, symmetric co-occurrence matrix, and

factor map) the minimum is one-field. For cross-correlation matrix or map and for asymmetric co-occurrence matrix the minimum is two-fields.



Using these virtual fields ...

- The user can create the same analytical views and perform most of the analytical steps (e.g. group operations, cleanup, thesaurus) as with the source datasets.
- These 'cross datasets' may be saved and reopened just like a regular VPT file. When a cross-dataset is reopened, TechOASIS checks to see if the source datasets for the cross-dataset are opened, and if not, those files will be opened automatically.

Note: Please refer to the following document(s) for full descriptions of these capabilities:

- SPS Document (CROSS-DATASET ANALYSIS – VIRTUAL FIELDS AND VIRTUAL DATASETS)

Update Datasets

The analytical products of TechOASIS are based on a "snapshot" of data. If the underlying data are found to be inaccurate or incomplete and need to be updated, the analytical products most likely need to change as well. The changes in the underlying data may have little or no effect, or they may have a substantial impact. The most straightforward way to understand the impact is to recreate the analytical products using the new updated data.

This activity in the project was to determine a way to streamline the process of recreating the analytical products when the underlying data is changed. The approaches considered included (a) a “recorder” that stored the analyst’s steps in a transaction list for playback against the updated dataset, and (b) a “replicator” that mimicked the analytical views from one dataset in another. These two approaches may be contrasted as “repeat the process” approach vs. a “copy the product” approach.

Both approaches were carried into a conceptual design stage, but neither proved to be entirely satisfactory. Both suffer from an inability to capture in a meaningful way the numerous decisions that a TechOASIS user makes during an analysis – for example cleanup confirmation, grouping, cut-off points and parameters for mapping. The membership in a group can easily be copied, but the decision to include or not include a new item from the new or updated records requires the analyst’s judgment.

After extensive review of the alternatives, the following approach was developed:

1. Establish a practice of archiving (a) the raw data and (b) the TechOASIS datasets after the initial import. These serve as the starting point when new data are introduced.
2. Save all cleanup activities and all manual-grouping activities as thesauri, and archive them with the raw datasets.
3. Carefully encode the analytical processes in automated scripts. A script command for importing raw data using default settings was developed to support this step (App.ImportFile).
4. When new or updated data arrive, import the data using the original import filter. If the filter cannot be located, it may be extracted from the initial TechOASIS dataset archived in step 1. If the import filter has been modified since the initial import, then import both the original raw data and the updated data using the modified import filter.
5. Use Data Fusion and/or Frankenrecords to merge the archived datasets with the updated dataset(s).
6. Use Remove Duplicates or Combine Duplicates to match, remove, or augment duplicates using multi-field matching.
7. Apply the Clean-up thesauri to the appropriate field(s), and then run cleanup. Merge any new cleanup decisions into the appropriate Cleanup thesaurus.
8. Apply the thesauri for replicating manually created groups. Review the groups to see if new items should be added to the groups, and merge any changes into the appropriate grouping thesaurus.
9. Finally, run the automated scripts to produce the analytical products using the updated dataset.

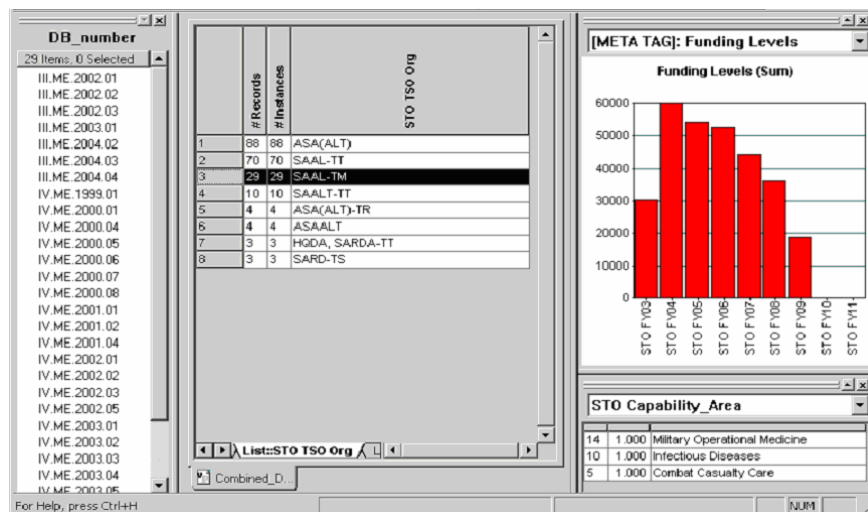
Note: Please refer to the following document(s) for full descriptions of these capabilities:

- VantagePoint User's Guide
- VantagePoint On-line help (Automation and Scripts / VantagePoint Scriptwriter Reference)

Graphical Display of Numeric Data

While TechOASIS relies heavily on numeric computation, the underlying data is usually integers and more specifically the number of records that meet a specific criterion (contains X, as in a list; or contains both X and Y, as in a co-occurrence matrix). Furthermore, typical “X” and “Y” values are text strings that are typically words, phrases, or indexing codes. Increasingly, TechOASIS is being called upon to also work with data where the text represents numbers (e.g., funding levels). While TechOASIS can parse and present these numbers, the presentation is limited to, for example, “the number of records that contain a funding level of \$1,000,000” when the desired information may be “the sum of the funding levels for the selected records is \$5,000,000”.

This project developed the capability to display numeric data (sums, means, max, min, and cumulative sums) in a detail window based on data contained in all fields with a common Metatag. The following illustration shows a display of the sum of the funding levels by FY for all STOs where the STO TSO Org = SAAL-TM. Each of the “STO FY0x” items in the detail window represents a field in the dataset. For example, the value for “STO FY03” (approximately 30000) is the sum of all values across all of SAAL-TM's records (segments) in the field “STO FY03”. That field (and all of the other “STO FY 0x” fields) has a Metatag set to “Funding Level”.



Note: Please refer to the following document(s) for full descriptions of these capabilities:

- VantagePoint User's Guide (Detail Window – Meta Tag Pop-up Menu)
- SPS Document (VISUALIZATION OF NUMERIC DATA)

WebQL interface

WebQL Enterprise from QL2 (www.ql2.com) software has an application programming interface (API) for various languages (including C++) that allows other software to run WebQL queries and collect the resulting output. The WebQL interface is embedded in TechOASIS allowing the user to invoke WebQL queries from some views in TechOASIS and automatically import the results into a new TechOASIS dataset. TechOASIS can run a query either locally or on a WebQL server hosted locally or elsewhere on the Internet.

As implemented in this program, the interface between TechOASIS and WebQL enables the user to:

- Run predefined WebQL queries from right-click menu pop-up in List and Detail Windows
- Optionally set WebQL parameters when running queries
- Automatically import query results into TechOASIS
- Automatically set data types for 4-digit years, numbers, and URL's (as 'file' datatype).

Part of the original requirements called for automatically applying the NLP parser to free text fields. During concept development and evaluation, feedback from the user community indicated that automatic import of NLP fields was not desirable. Their assessment was that in this context prior to NLP import, the user should define a 'protected' list of terms for the NLP parser (see also the SPS Document section ENTITY EXTRACTION WITH NLP PARSING). An automatically imported list without this 'protection' will slow the import process, and the automatically imported field most likely be discarded in favor of a more targeted approach.

Note: Please refer to the following document(s) for full descriptions of these capabilities:

- SPS Document (WEBQL INTERFACE)

Link Analysis

In the context of this project, link analysis is a technique we are tasked to review and implement. Link analysis tries to establish a relationship between items even if they do not appear frequently (or at all) in the same documents. Link analysis might also be used to determine if two documents are related even though they share few or no terms. This technique should also be able to determine if an item and a document are related even if the item does not appear in that document. A main goal is to be able to associate items and documents across dissimilar datasets (e.g. datasets with disparate source material like patents versus journal publications).

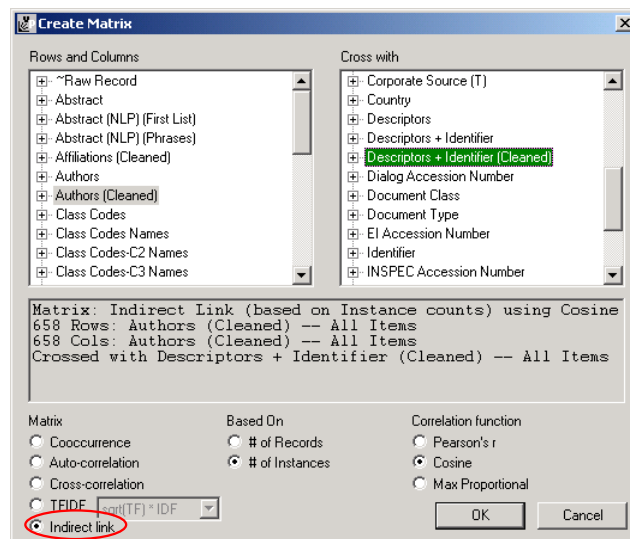
The SPS Document describes a novel measure for link analysis called 'indirect link correlation', and its implementation within TechOASIS.

The following sections describe new tools to support link analysis, including a new indirect link matrix type and a new type of matrix view that makes link analysis

practicable. It also includes new functionality in the detail windows that shows which terms contribute (or not) to a relationship between terms. Finally, this section relates a complete example of using the components described here for a link analysis task.

Indirect Link Matrix

To investigate the indirect link correlation measure, we have implemented a new matrix type in TechOASIS (see screen shot below). The indirect link matrix is created using the same dialog as all other matrix types in TechOASIS. Note that you can choose the same parameters – “Based On” and “Correlation function” from the indirect link matrix as you can with an auto-correlation or cross-correlation matrix.



The resulting indirect link matrix looks like this:

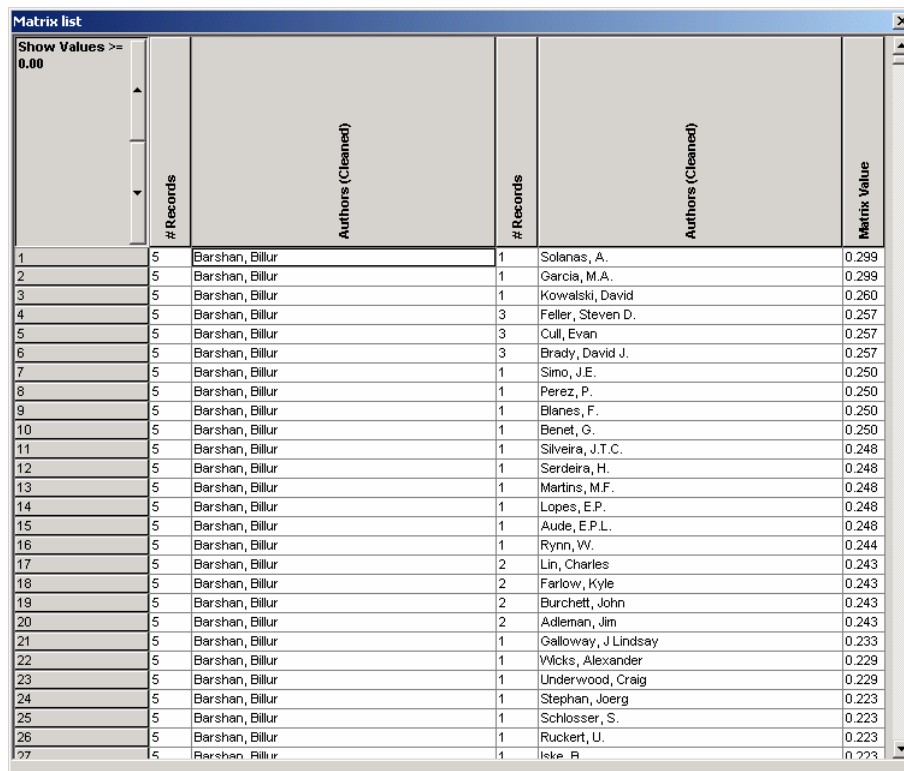
Show Values >= 0.00											
Indirect Link Crossed With: Descriptors + Identifier (Cleaned) # of Records Cosine											
	Barshan, Billur	Aytac, Tayfun	Henning, P.F.	Chadha, Suneet	Plaza, J.	Rodrigo, M.T.	Rodriguez, P.	Rosenbury, Tom	Sanchez, F.J.	Cull, Evan	Barbael Hosain
Barshan, Billur	0.000	0.034	0.069	0.069	0.086	0.086	0.086	0.033	0.086	0.258	0.000
Aytac, Tayfun	0.034	0.000	0.057	0.057	0.071	0.071	0.071	0.038	0.071	0.256	0.000
Henning, P.F.	0.069	0.057	0.000	0.000	0.082	0.082	0.082	0.022	0.082	0.094	0.000
Chadha, Suneet	0.069	0.057	0.000	0.000	0.082	0.082	0.082	0.022	0.082	0.094	0.000
Plaza, J.	0.086	0.071	0.082	0.082	0.000	0.000	0.000	0.014	0.000	0.078	0.000
Rodrigo, M.T.	0.086	0.071	0.082	0.082	0.000	0.000	0.000	0.014	0.000	0.078	0.000
Rodriguez, P.	0.086	0.071	0.082	0.082	0.000	0.000	0.000	0.014	0.000	0.078	0.000
Rosenbury, Tom	0.033	0.038	0.022	0.022	0.014	0.014	0.014	0.000	0.014	0.066	0.000
Sanchez, F.J.	0.086	0.071	0.082	0.082	0.000	0.000	0.000	0.014	0.000	0.078	0.000
Cull, Evan	0.258	0.256	0.094	0.094	0.078	0.078	0.078	0.066	0.078	0.000	0.000
Barbael Hosain	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.031	0.000
Holland, Stephen K.	0.053	0.041	0.225	0.225	0.074	0.074	0.074	0.000	0.074	0.070	0.000
Lee, Seon-Woo	0.075	0.074	0.029	0.029	0.071	0.071	0.071	0.036	0.071	0.085	0.000
Sumpter, Neil	0.218	0.207	0.120	0.120	0.170	0.170	0.170	0.043	0.170	0.264	0.000
Lee, Sooyong	0.210	0.207	0.040	0.040	0.099	0.099	0.099	0.034	0.099	0.142	0.000
Diezhandino, J.	0.086	0.071	0.082	0.082	0.000	0.000	0.000	0.014	0.000	0.078	0.000
Todoroki, Akira	0.000	0.000	0.000	0.000	0.038	0.038	0.038	0.052	0.038	0.018	0.000
Torquemada, Md.C.	0.086	0.071	0.082	0.082	0.000	0.000	0.000	0.014	0.000	0.078	0.000

Note that this is pretty much a jumble of values with no real ordering apparent. It's not like a co-occurrence matrix where large values tend to appear toward the top left corner of the matrix.

So, how to find the interesting, large values in this matrix? Searching, sorting, and/or flooding is too tedious and macros perhaps too inefficient. Copying and pasting into Excel is not really an option since Excel can only hold rather modest-sized matrices.

Matrix List

To address the difficulties of finding high values in a large matrix with no real apparent ordering, we have developed a new kind of view called the 'matrix list'. The matrix list is currently implemented in a dialog but it might be included as another TechOASIS main view in the future. The matrix list is a list that has a row for each cell in the matrix. So, for the above matrix the matrix list looks like this:



	# Records	Authors (Cleaned)	# Records	Authors (Cleaned)	Matrix Value
1	5	Barshan, Billur	1	Solanas, A.	0.299
2	5	Barshan, Billur	1	Garcia, M.A.	0.299
3	5	Barshan, Billur	1	Kowalski, David	0.260
4	5	Barshan, Billur	3	Feller, Steven D.	0.257
5	5	Barshan, Billur	3	Cull, Evan	0.257
6	5	Barshan, Billur	3	Brady, David J.	0.257
7	5	Barshan, Billur	1	Simo, J.E.	0.250
8	5	Barshan, Billur	1	Perez, P.	0.250
9	5	Barshan, Billur	1	Blanes, F.	0.250
10	5	Barshan, Billur	1	Benet, G.	0.250
11	5	Barshan, Billur	1	Silveira, J.T.C.	0.248
12	5	Barshan, Billur	1	Serdeira, H.	0.248
13	5	Barshan, Billur	1	Martins, M.F.	0.248
14	5	Barshan, Billur	1	Lopes, E.P.	0.248
15	5	Barshan, Billur	1	Aude, E.P.L.	0.248
16	5	Barshan, Billur	1	Rynn, W.	0.244
17	5	Barshan, Billur	2	Lin, Charles	0.243
18	5	Barshan, Billur	2	Farlow, Kyle	0.243
19	5	Barshan, Billur	2	Burchett, John	0.243
20	5	Barshan, Billur	2	Adleman, Jim	0.243
21	5	Barshan, Billur	1	Galloway, J Lindsay	0.233
22	5	Barshan, Billur	1	Wicks, Alexander	0.229
23	5	Barshan, Billur	1	Underwood, Craig	0.229
24	5	Barshan, Billur	1	Stephan, Joerg	0.223
25	5	Barshan, Billur	1	Schlosser, S.	0.223
26	5	Barshan, Billur	1	Ruckert, U.	0.223
27	5	Barshan, Billur	1	Iske, R.	0.223

The left-most column has a row-count. The next column shows the number of records for the items in the rows of the matrix. The third column from the left is the actual item label for the rows of the matrix. The next two columns shows the same information but for the items in the matrix columns instead of rows. The last column has the matrix value – this is the value that is found in the intersecting cell of the row indicated by the row item and the column indicated by the column item.

The matrix list can be generated for any type of matrix (e.g. co-occurrence, TFIDF, etc.), not just the indirect link matrix. The matrix list can currently be accessed from the ‘Tools’ menu – “List cells in matrix ...”.

The matrix list interacts with its underlying matrix view when the user selects items in the matrix list. When the user selects a cell in a matrix list in the second column (row item # records) or the third column (row item label) the entire row corresponding to this row item is selected in the underlying matrix view. This interaction between the matrix list and the underlying matrix behaves similarly for column items. If the user selects a cell in the Matrix Value column of the matrix list the single cell corresponding to this value in the underlying matrix is selected. Any cells in the matrix list can be multi-selected resulting in multiple selection of rows, columns, and/or cells in the underlying matrix. The following two screen shots illustrate this:

Author (Cleaned)	# Records	Descriptors (Cleaned) # of Records Max Proportional	Keller, Paul	Lin, Willie	Elbert, Jim	Takiguchi, Jun-ichi	Iwama, Kiyoshi	Sugie, Hiroshi	Kato, Masanori	Kiyonaga, Takayuki	Hashizume, Takumi	Inoue, Fumihiro	Yoshino, Kyoji	Onote, Yutaro	Kamada, Toru	Okawa, Koichi	Page, Lance	Lutzeler, M.	Arai, Y.	Fuji, T.	Pearson, C.M.	Enbert, P.I.
1	12	Hebert, Martial	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.857	0.833	0.833	0.833	0.833	0.8
2	2	Page, Lance	1.000	0.667	0.667	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.333	0.333	0.000	0.000	0.333	0.333	0.333	0.3
3	3	Shen, C. N.	1.000	0.667	0.667	0.250	0.250	0.25														
4	3	Langer, D.	1.000	1.000	1.000	0.750	0.750	0.75														
5	2	Kurtz, Jay	0.500	0.667	0.667	0.750	0.750	0.75														
6	1	Tran, T.T.	0.500	0.333	0.333	0.500	0.500	0.50														
7	1	Feder, Hans Jacob S.	0.500	0.333	0.333	0.500	0.500	0.50														
8	3	White, Mark W.	0.500	0.333	0.333	0.500	0.500	0.50														
9	1	Bachnak, Rafic	0.500	0.333	0.333	0.250	0.250	0.25														
10	2	Sutton, John C. III	0.500	0.333	0.333	0.500	0.500	0.50														
11	3	Hague, T.	0.500	0.333	0.333	0.250	0.250	0.25														
12	1	Leonard, John J.	0.500	0.333	0.333	0.500	0.500	0.50														
13	1	Wang, Yanxing	0.500	0.667	0.667	0.250	0.250	0.25														
14	2	Banta, Larry	0.500	0.667	0.667	0.250	0.250	0.25														
15	1	Rawson, Keith D.	0.500	0.667	0.667	0.250	0.250	0.25														
16	1	Smith, Chris M.	0.500	0.333	0.333	0.500	0.500	0.50														
17	3	Desai, Rajiv	0.500	0.667	0.667	0.250	0.250	0.25														
18	2	Loch, John	0.500	0.667	0.667	0.250	0.250	0.25														
19	3	Miller, David P.	0.500	0.667	0.667	0.250	0.250	0.25														
20	1	Wettergreen, David	0.500	0.667	0.667	0.250	0.250	0.25														
21	1	Okamoto, Hironobu	0.500	0.333	0.333	0.250	0.250	0.25														
22	1	Izumi, Tetsujiro	0.500	0.333	0.333	0.250	0.250	0.25														
23	1	Kishi, Hiroo	0.500	0.333	0.333	0.250	0.250	0.25														
24	1	Cordeuener, L.	0.500	0.333	0.333	0.250	0.250	0.25														

Author (Cleaned)	# Records	Indirect Link Crossed With: Descriptors (Cleaned) # of Records Max Proportional	Kelly, Alonzo	Tsugawa, Sadayuki	Stentz, Anthony	Malik, Jitendra	Luo, Ren C.	Distante, A.	Kehtarnavaz, Nasser D.	Whyte, Hugh-Durrant	Gat, Erann	Janet, Jason A.	Griswold, Norman C.	Hebert, Martial	Kehtarnavaz, Nasser	Mathies, Larry	Marchant, J.A.	Meystel, Alex	Anon (Ed.)
13	Anon (Ed.)	0.542	0.513	0.500	0.489	0.487	0.450	0.432	0.422	0.410	0.396	0.393	0.366	0.366	0.355	0.282	0.158	0.000	
12	Hebert, Martial	0.313	0.333	0.235	0.234	0.282	0.225	0.341	0.400	0.333	0.302	0.286	0.000	0.244	0.387	0.282	0.053	0.366	
10	Stentz, Anthony	-0.020	0.333	0.000	0														
9	Kehtarnavaz, Nasser	0.195	0.179	0.195	0														
8	Kelly, Alonzo	0.000	0.282	-0.020	0														
7	Whyte, Hugh-Durrant	0.444	0.179	0.556	0														
7	Janet, Jason A.	0.250	0.103	0.302	0														
6	Tsugawa, Sadayuki	0.282	0.000	0.333	0														
6	Griswold, Norman C.	0.179	0.214	0.179	0														
5	Distante, A.	0.250	0.077	0.325	0														
5	Meystel, Alex	0.105	0.105	0.105	0														
5	Gat, Erann	0.385	0.051	0.410	0														
5	Marchant, J.A.	0.256	0.154	0.282	0														
5	Malik, Jitendra	0.234	0.359	0.213	0														
5	Kehtarnavaz, Nasser D.	0.227	0.231	0.250	0														
5	Mathies, Larry	0.187	0.258	0.258	0														
5	Luo, Ren C.	0.282	0.103	0.333	0														

Author (Cleaned)	# Records	Author (Cleaned)	# Records	Author (Cleaned)	Matrix Value
1	10	Stentz, Anthony	7	Whyte, Hugh-Durrant	0.555
2	7	Whyte, Hugh-Durrant	10	Stentz, Anthony	0.555
3	13	Anon (Ed.)	8	Kelly, Alonzo	0.541
4	8	Kelly, Alonzo	13	Anon (Ed.)	0.541
5	13	Anon (Ed.)	6	Tsugawa, Sadayuki	0.512
6	6	Tsugawa, Sadayuki	13	Anon (Ed.)	0.512
7	13	Anon (Ed.)	10	Stentz, Anthony	0.5
8	10	Stentz, Anthony	13	Anon (Ed.)	0.5

Link Analysis Enhancements to the Detail Window

The selection interaction between the matrix list and the underlying matrix view are critical for identifying the terms that both do and do not overlap between the two items being compared for link analysis. By viewing the overlapping and non-overlapping terms the user may be able to infer the ‘nature’ of the relationship between two indirectly linked items.

When an *entire* row and/or entire column is selected in a matrix view, items in the detail views are shaded with different colors to indicate if the item corresponds with a row item selected in the matrix view, a column item selection in the matrix view, or both.

Look at the following screen shot showing an Authors X Authors indirect link matrix with various colors in the detail view:

Matrix (Cleaned)		# Records									
		2	2	1	1	1	1	1	1	1	1
Show Values >= 0.00											
Indirect Link Crossed With:											
Descriptors (Cleaned)											
# of Records											
Max Proportional											
		Mandelbaum, R.	Pagac, Daniel	Lacaze, A.	Murphy, K.	Balija, Shumeet	Lambole, M.	Proy, C.	Rastel, L.	Naumen Trono, T.	
1	Faugeras, O.D.	0.333	0.222	0.286	0.286	0.571	0.143	0.143	0.143	0.14	
1	Audren, J.T.	0.333	0.222	0.286	0.286	0.571	0.143	0.143	0.143	0.14	
5	Shirai, Yoshiaki	0.438	0.818	0.571	0.571	0.571	0.429	0.429	0.429	0.42	
1	Schudel, D.S.	0.167	0.333	0.167	0.167	0.333	0.167	0.167	0.167	0.16	
1	Bornstein, Jonathan A.	0.750	0.250	0.571	0.571	0.143	0.429	0.429	0.429	0.42	
1	England, A.G.	0.167	0.333	0.167	0.167	0.333	0.167	0.167	0.167	0.16	
1	Snyder, W.E.	0.167	0.333	0.167	0.167	0.333	0.167	0.167	0.167	0.16	
1	Chamnonthai, Kosin	0.333	0.333	0.333	0.333	0.333	0.333	0.333	0.333	0.33	
1	Ozawa, Shinji	0.333	0.333	0.333	0.333	0.333	0.333	0.333	0.333	0.33	
1	Wang, Ling-Ling	0.400	0.200	0.286	0.286	0.429	0.286	0.286	0.286	0.28	
1	Ku, Pao-Yu	0.400	0.200	0.286	0.286	0.429	0.286	0.286	0.286	0.28	
5	Distante, A.	0.188	0.364	0.429	0.429	0.571	0.286	0.286	0.286	0.28	
2	Maurer, M.	0.250	0.182	0.143	0.143	0.286	0.143	0.143	0.143	0.14	
1	Tsai, Wen-Hsiang	0.400	0.200	0.286	0.286	0.429	0.286	0.286	0.286	0.28	
1	Williams, J.H.	0.333	0.167	0.167	0.167	0.167	0.167	0.167	0.167	0.16	
1	Rayment, P.J.	0.333	0.167	0.167	0.167	0.167	0.167	0.167	0.167	0.16	
1	Lallement, A.	0.500	0.125	0.286	0.286	0.286	0.286	0.286	0.286	0.28	
5	Nebot, Eduardo Mario	0.188	0.000	0.286	0.286	0.286	0.286	0.286	0.286	0.28	
2	Hahad, Tienhu	0.231	0.001	0.286	0.286	0.286	0.429	0.429	0.429	0.42	

In the above matrix the entire row for the author “Shirai, Yoshiaki” and the entire column for author “Pagac, Daniel” is selected. The “Descriptors (Cleaned)” field was used to compute the indirect link correlations and is displayed in the detail window. The Descriptors that co-occur with both the authors (i.e. the items that co-occur with all the items for completely selected rows and columns) appear shaded with green. For link analysis this corresponds to ‘overlapping’ terms. Red-shaded items are Descriptors that co-occur with the selected row items (i.e. in this case “Shirai, Yoshiaki”) that *do not* co-occur with the selected column items. Similarly, yellow-shaded items are those that co-occur with the selected column items (i.e. in this case “Pagac, Daniel”) that *do not* co-occur with the selected row items.

Link Analysis Example

This section describes a complete example of using the indirect link correlation measure, the matrix list, and the coloring in the detail view in a link analysis task.

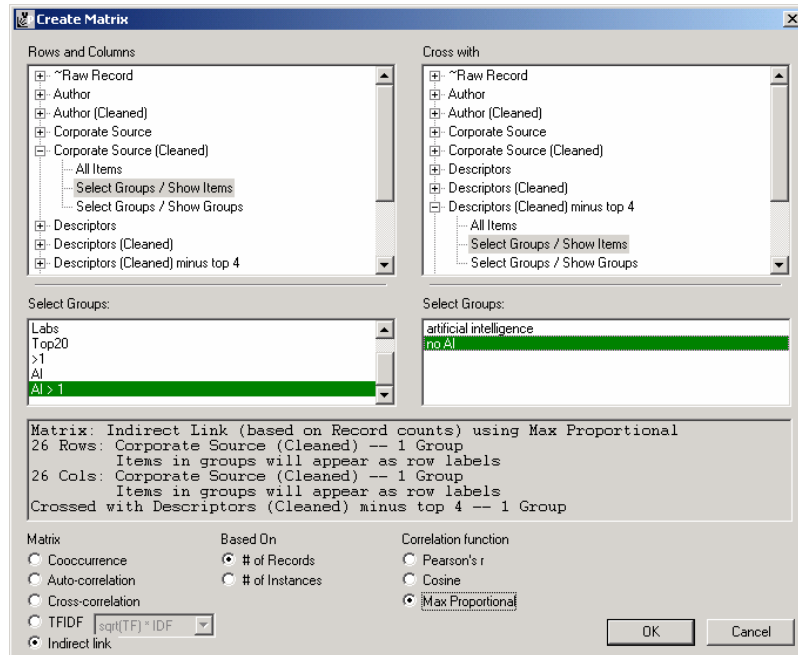
Suppose an analyst was interested in the topic of ‘automatic navigation’, particularly which institutions were publishing about ‘artificial intelligence’ in this topic area. One of the questions he wants to answer is ‘of the institutions using artificial intelligence in their research concerning automatic navigation, which are doing similar things but not collaborating?’

After obtaining an EI Compendex dataset on automatic navigation he does a search on “artificial intelligence” in the Descriptors field of this dataset and groups the Corporate Source items that co-occur with the found items. He decides he is only interested in the institutions with more than one publication in this dataset. Following is a list of Corporate Sources that co-occur with “artificial intelligence” with greater than one record in this dataset:

Charles Stark Draper Lab, Cambridge, MA, USA	Sch of Inf & Comput Sci, Georgia Inst of Technol, Atlanta, GA, USA
Defence Research Agency, Engl	SONATECH Inc
DLR, Ger	Syracuse Univ, Syracuse, NY, USA
Heriot-Watt Univ, Intelligent Autom Lab, Edinburgh, Scotl	Texas Transp Inst, Texas A&M Univ, Texas, TX, USA
Jet Propulsion Lab., Pasadena, CA, USA	The Pennsylvania State Univ, University Park, PA, USA
Lockheed Missiles & Space Co, Sunnyvale, CA, USA	U.S. Bureau of Mines, Pittsburgh, PA, USA
Martin Marietta Astronautics Group, Denver, CO, USA	Univ of California, Riverside, CA, USA
Mazda Motor Corp, Jpn	Univ of Darmstadt, Darmstadt, Ger
Navy Center for Applied Research in Artificial Intelligence	Univ of Girona, Catalonia, Spain
Oak Ridge Natl Lab, TN, USA	Univ of Southampton, Highfield, Engl
Office of Natl d'Etudes et de Recherche Aerospatiales, Toulouse, Fr	Univ of Technology of Compiegne, Compiegne, Fr
Osaka Univ, Osaka, Jpn	Univ. of Pennsylvania, Philadelphia, PA, USA
Santa Clara Univ, Santa Clara, CA, USA	Universitaet der Bundeswehr, Neubiberg, Ger

These Corporate Source items are added to a new group called “AI > 1”.

He then calculates the indirect link measure between each of these institutions based on cleaned Descriptors. Since the top four terms in the Descriptors field are quite prevalent he eliminates these from the matrix calculation. He also eliminates the terms that match “artificial intelligence” since all the institutions above co-occur with these terms. The indirect link calculation is based on the Max Proportional correlation function:



The resulting matrix looks like this (note the jumble of numbers):

VantagePoint [auto navigation.vpl]																	
File Edit View Sheets Fields Groups Tools Scripts Window Help																	
Title																	
0 Items, 0 Selected																	
Reset Corporate Source (Cleaned)																	
# Records		# Records															
Show Values >= 0.00																	
Indirect Link Crossed With Descriptors (Cleaned) minus top 4 Groups (Items) # of Records Max Proportional																	
Corporate Source (Cleaned)																	
# Records																	
1	19	Univ of California, Riverside, CA, USA	0.000	0.373	0.609	0.250	0.529	0.462	0.457	0.250	0.519	0.375	0.524				
2	16	Texas Transp Inst, Texas A&M Univ, Texas	0.273	0.006	0.326	0.321	0.042	0.615	0.371	0.250	0.407	0.500	0.333				
3	9	The Pennsylvania State Univ, University Park	0.059	0.326	0.000	0.143	0.305	0.385	0.314	0.143	0.222	0.188	0.238				
4	9	Santa Clara Univ, Santa Clara, CA, USA	0.250	0.321	0.143	0.000	0.059	0.308	0.143	0.071	0.111	0.313	0.180				
5	8	Jet Propulsion Lab, Pasadena, CA, USA	0.029	0.412	0.363	0.059	0.000	0.231	0.118	0.235	0.176	0.000	0.235				
6	8	Oak Ridge Natl Lab, TN, USA	0.462	0.616	0.385	0.308	0.231	0.000	0.154	0.077	0.231	0.231	0.231				
7	8	Universität der Bundeswehr, Neuburg	0.457	0.371	0.314	0.143	0.118	0.154	0.000	0.179	0.222	0.250	0.238				
8	6	Martin Marietta Astronautics Group, Denver	0.250	0.250	0.143	0.071	0.235	0.077	0.179	0.000	0.250	0.063	0.180				
9	6	Sch of Int & Comput Sci, Georgia Inst of Tech	0.118	0.176	0.231	0.111	0.176	0.231	0.231	0.250	0.000	0.250	0.180				
10	5	Univ of Pennsylvania, Philadelphia, PA, US	0.039	0.500	0.168	0.313	0.000	0.231	0.059	0.063	0.250	0.000	0.125				
11	4	Charles Stark Draper Lab, Cambridge, MA	0.524	0.333	0.238	0.190	0.235	0.231	0.238	0.190	0.190	0.125	0.000				
12	4	Herod-Watt Univ, Intelligent Autom Lab, Ed	0.039	0.267	0.200	0.067	0.133	0.077	0.200	0.200	0.200	0.067	0.267				
13	3	Lockheed Missiles & Space Co, Sunnyvale	0.500	0.667	0.333	0.333	0.500	0.333	0.167	0.333	0.500	0.167	0.333				
14	3	Univ of Technology of Comptec, Paris, France	0.068	0.538	0.368	0.154	0.077	0.077	0.368	0.154	0.231	0.154	0.231				
15	3	Syracuse Univ, Syracuse, NY, USA	0.000	0.400	0.000	0.400	0.000	0.600	0.000	0.500	0	0.000	0.500				
16	2	Office of Natl Etudes et de Recherche en Ar	0.692	0.231	0.231	0.077	0.154	0.077	0.462	0.000	0.308	0.000	0.231				
17	2	Defence Research Agency, Engl	0.625	0.625	0.250	0.375	0.125	0.250	0.500	0.125	0.375	0.000	0.250				
18	2	Univ of Southampton, Hightest, Engl	0.667	0.667	0.333	0.167	0.333	0.167	0.167	0.667	0.000	0.333					
19	2	Crata Univ, Crata, Jpn	0.615	0.615	0.365	0.154	0.154	0.154	0.365	0.154	0.154	0.365					
20	2	Mazda Motor Corp, Jpn	1.000	1.000	0.500	0.500	0.000	0.500	0.500	0.000	0.500	0.000	0.500				
21	2	Navy Center for Applied Research in Artifi	0.636	0.456	0.273	0.182	0.273	0.182	0.364	0.273	0.182	0.091	0.456				
22	2	U.S. Bureau of Mines, Pittsburgh, PA, US	0.000	0.667	0.333	0.000	0.667	0.333	0.000	0.000	0.667	0.000	0.667				
23	2	MINATECH Inc	0.000	0.500	0.500	0.500	0.500	0.000	0.000	0.500	0.000	0.500					
24	2	DLR, Ger	0.636	0.545	0.273	0.091	0.182	0.182	0.182	0.364	0.273	0.091	0.091				
25	2	Univ of Ghent, Catalonia, Spain	0.500	0.083	0.417	0.000	0.000	0.000	0.167	0.083	0.167	0.000	0.000				
26	2	Research Center, Research, Jpn	0.462	0.462	0.231	0.231	0.231	0.231	0.231	0.231	0.231	0.231	0.231				
Matrix Corporate Source (Cleaned)(Items)XCorporate																	

To make some sense out of this matrix he then creates a matrix list (Tools menu > List Cells in Matrix...). Since he is interested in high indirect link values, he floods the

matrix to 0.50. He is also mainly interested in the top publishing institution so he sorts the row items by # Records and gets the following matrix list:

Matrix list				
Show Values >= 0.50				
	# Records	Corporate Source (Cleaned)	# Records	Corporate Source (Cleaned)
1	19	Univ of California, Riverside, CA, USA	2	SONATECH Inc
2	19	Univ of California, Riverside, CA, USA	2	Mazda Motor Corp, Jpn
3	19	Univ of California, Riverside, CA, USA	3	Univ of Technology of Compiègne, Compiègne, Fr
4	19	Univ of California, Riverside, CA, USA	2	Office of Natl d'Etudes et de Recherche Aérospatiale
5	19	Univ of California, Riverside, CA, USA	2	Univ of Southampton, Highfield, Engl
6	19	Univ of California, Riverside, CA, USA	2	U.S. Bureau of Mines, Pittsburgh, PA, USA
7	19	Univ of California, Riverside, CA, USA	2	Univ of Darmstadt, Darmstadt, Ger
8	19	Univ of California, Riverside, CA, USA	2	Navy Center for Applied Research in Artificial Intell
9	19	Univ of California, Riverside, CA, USA	2	DLR, Ger
10	19	Univ of California, Riverside, CA, USA	2	Defence Research Agency, Engl
11	19	Univ of California, Riverside, CA, USA	2	Osaka Univ, Osaka, Jpn
12	19	Univ of California, Riverside, CA, USA	9	The Pennsylvania State Univ, University Park, PA, U
13	19	Univ of California, Riverside, CA, USA	3	Heriot-Watt Univ, Intelligent Autom Lab, Edinburgh, S
14	19	Univ of California, Riverside, CA, USA	6	Jet Propulsion Lab., Pasadena, CA, USA
15	19	Univ of California, Riverside, CA, USA	4	Charles Stark Draper Lab, Cambridge, MA, USA
16	19	Univ of California, Riverside, CA, USA	6	Sch of Inf & Comput Sci, Georgia Inst of Technol, At
17	16	Texas Transp Inst, Texas A&M Univ, Texas, TX, US	2	SONATECH Inc
18	16	Texas Transp Inst, Texas A&M Univ, Texas, TX, US	2	Mazda Motor Corp, Jpn
19	16	Texas Transp Inst, Texas A&M Univ, Texas, TX, US	3	Lockheed Missiles & Space Co, Sunnyvale, CA, US
20	16	Texas Transp Inst, Texas A&M Univ, Texas, TX, US	2	Univ of Southampton, Highfield, Engl
21	16	Texas Transp Inst, Texas A&M Univ, Texas, TX, US	2	U.S. Bureau of Mines, Pittsburgh, PA, USA
22	16	Texas Transp Inst, Texas A&M Univ, Texas, TX, US	2	Univ of Darmstadt, Darmstadt, Ger
23	16	Texas Transp Inst, Texas A&M Univ, Texas, TX, US	2	Defence Research Agency, Engl
24	16	Texas Transp Inst, Texas A&M Univ, Texas, TX, US	6	Oak Ridge Natl Lab, TN, USA
25	16	Texas Transp Inst, Texas A&M Univ, Texas, TX, US	2	Osaka Univ, Osaka, Jpn
26	16	Texas Transp Inst, Texas A&M Univ, Texas, TX, US	2	DLR, Ger

He focuses on two top publishers: Univ of California, Riverside, CA, USA (19 records) and The Pennsylvania State Univ, University Park, PA, USA (9 records) with an indirect link value of 0.608. *Note that these institutions share no records.* He selects both items in the matrix list:

Matrix list				
Show Values >= 0.50				
	# Records	Corporate Source (Cleaned)	# Records	Corporate Source (Cleaned)
1	19	Univ of California, Riverside, CA, USA	2	SONATECH Inc
2	19	Univ of California, Riverside, CA, USA	2	Mazda Motor Corp, Jpn
3	19	Univ of California, Riverside, CA, USA	3	Univ of Technology of Compiègne, Compiègne, Fr
4	19	Univ of California, Riverside, CA, USA	2	Office of Natl d'Etudes et de Recherche Aérospatiale
5	19	Univ of California, Riverside, CA, USA	2	Univ of Southampton, Highfield, Engl
6	19	Univ of California, Riverside, CA, USA	2	U.S. Bureau of Mines, Pittsburgh, PA, USA
7	19	Univ of California, Riverside, CA, USA	2	Univ of Darmstadt, Darmstadt, Ger
8	19	Univ of California, Riverside, CA, USA	2	Navy Center for Applied Research in Artificial Intell
9	19	Univ of California, Riverside, CA, USA	2	DLR, Ger
10	19	Univ of California, Riverside, CA, USA	2	Defence Research Agency, Engl
11	19	Univ of California, Riverside, CA, USA	2	Osaka Univ, Osaka, Jpn
12	19	Univ of California, Riverside, CA, USA	9	The Pennsylvania State Univ, University Park, PA, U
13	19	Univ of California, Riverside, CA, USA	3	Heriot-Watt Univ, Intelligent Autom Lab, Edinburgh, S
14	19	Univ of California, Riverside, CA, USA	6	Jet Propulsion Lab., Pasadena, CA, USA
15	19	Univ of California, Riverside, CA, USA	4	Charles Stark Draper Lab, Cambridge, MA, USA
16	19	Univ of California, Riverside, CA, USA	6	Sch of Inf & Comput Sci, Georgia Inst of Technol, At
17	16	Texas Transp Inst, Texas A&M Univ, Texas, TX, US	2	SONATECH Inc
18	16	Texas Transp Inst, Texas A&M Univ, Texas, TX, US	2	Mazda Motor Corp, Jpn
19	16	Texas Transp Inst, Texas A&M Univ, Texas, TX, US	3	Lockheed Missiles & Space Co, Sunnyvale, CA, US
20	16	Texas Transp Inst, Texas A&M Univ, Texas, TX, US	2	Univ of Southampton, Highfield, Engl
21	16	Texas Transp Inst, Texas A&M Univ, Texas, TX, US	2	U.S. Bureau of Mines, Pittsburgh, PA, USA
22	16	Texas Transp Inst, Texas A&M Univ, Texas, TX, US	2	Univ of Darmstadt, Darmstadt, Ger
23	16	Texas Transp Inst, Texas A&M Univ, Texas, TX, US	2	Defence Research Agency, Engl
24	16	Texas Transp Inst, Texas A&M Univ, Texas, TX, US	6	Oak Ridge Natl Lab, TN, USA
25	16	Texas Transp Inst, Texas A&M Univ, Texas, TX, US	2	Osaka Univ, Osaka, Jpn
26	16	Texas Transp Inst, Texas A&M Univ, Texas, TX, US	2	DLR, Ger

The corresponding row and column are automatically selected in the underlying matrix view and the detail window is automatically updated:

VantagePoint - [auto-navigation.vpt]

File Edit View Sheets Fields Groups Tools Scripts Window Help

28 Items, 1 Selected

Accurate estimation of...
Analyses of underwater...
Automatic symbolic trait...
Autonomous fuzzy logic...
Autonomous navigation...
Autonomous path plan...
Conditional sequencing...
Detection of obstacles...
Differential GPS with lat...
Extended linear quadrat...
Fuzzy logic architecture...
Introduction to multisen...
Latency compensation...
Learning to perceive ob...
Linear quadratic Gauss...
Motion estimation via d...
Obstacle detection for...
Obstacle representation...
Path planning of an aut...
Practical autonomous p...
Rapid hover-to-forward...
Real-time 2-D feature d...
Robot navigation by co...
Rocky 7 rover: A Mas...
System for obstacle det...
Tracking for fully actual...
Vehicle longitudinal con...
Vision-based planetary r...

Corporate Source (Cleaned)

Reset Show Values >= 0.00 # Records 19

Indirect Link
Crossed With Descriptors
(Cleaned) minus top 4
1 Groups (Items)
Not Records
Max Proportional

# Records	Corporate Source (Cleaned)	1	2	3	4	5	6	7	8	9
1	Univ of California, Riverside, CA, USA	0.000	0.373	0.609	0.250	0.529	0.462	0.457	0.25	
2	Texas Transp Inst, Texas A&M Univ, Texa	0.273	0.000	0.325	0.321	0.412	0.615	0.371	0.25	
3	The Pennsylvania State Univ, University Pa	0.609	0.325	0.000	0.143	0.353	0.385	0.314	0.14	
4	Santa Clara Univ, Santa Clara, CA, USA	0.250	0.321	0.143	0.000	0.059	0.308	0.143	0.07	
5	Jet Propulsion Lab, Pasadena, CA, USA	0.529	0.412	0.353	0.059	0.000	0.231	0.118	0.23	
6	Oak Ridge Natl Lab, TN, USA	0.462	0.615	0.385	0.308	0.231	0.000	0.154	0.07	
7	Universität der Bundeswehr, Neuberg	0.457	0.371	0.314	0.143	0.118	0.154	0.080	0.17	
8	Martin Marietta Astronautics Group, Denv	0.250	0.250	0.143	0.071	0.225	0.077	0.175	0.00	
9	Sch of Inf & Comput Sci, Georgia Inst of Te	0.519	0.407	0.222	0.111	0.176	0.231	0.222	0.25	
10	Univ of Pennsylvania, Philadelphia, PA, US	0.375	0.500	0.188	0.313	0.000	0.231	0.250	0.06	
11	Charles Stark Draper Lab, Cambridge, MA	0.524	0.333	0.238	0.190	0.235	0.231	0.238	0.19	
12	Heriot-Watt Univ, Intelligent Autom Lab, Ed	0.533	0.257	0.200	0.067	0.133	0.077	0.200	0.20	
13	Lockheed Missiles & Space Co, Sunnyvale	0.500	0.667	0.333	0.333	0.500	0.333	0.167	0.33	
14	Univ of Technology of Compiègne, Compiè	0.692	0.536	0.308	0.154	0.077	0.077	0.308	0.15	
15	Syracuse Univ, Syracuse, NY, USA	0.000	0.400	0.000	0.400	0.000	0.600	0.000	0.00	
16	Office of Natl d'Etudes et de Recherche A	0.692	0.231	0.231	0.077	0.154	0.077	0.462	0.00	
17	Defence Research Agency, Engl	0.625	0.625	0.250	0.375	0.125	0.250	0.500	0.12	
18	Univ of Southampton, Highfield, Engl	0.667	0.667	0.333	0.167	0.333	0.333	0.167	0.16	
19	Osaka Univ, Osaka, Jpn	0.615	0.615	0.308	0.154	0.154	0.154	0.308	0.15	
20	Mazda Motor Corp, Jpn	1.000	1.000	0.500	0.500	0.000	0.000	0.500	0.00	
21	Navy Center for Applied Research in Artifi	0.636	0.455	0.273	0.182	0.273	0.182	0.364	0.27	
22	U.S. Bureau of Mines, Pittsburgh, PA, USA	0.667	0.667	0.333	0.000	0.667	0.333	0.000	0.00	
23	SONATECH Inc	1.000	1.000	0.500	0.500	0.500	0.500	0.000	0.00	
24	DLR, Ger	0.636	0.545	0.273	0.091	0.182	0.182	0.182	0.36	
25	Univ of Girona, Catalonia, Spain	0.500	0.083	0.417	0.000	0.000	0.000	0.167	0.08	
26	Univ of Research and Technol, Mosc	0.625	0.625	0.250	0.375	0.125	0.250	0.500	0.12	

Descriptors (Cleaned) minus top 4

2	Air navigation
2	Aircraft
1	Approximation theory
4	Artificial intelligence
2	Calibration
1	Cameras
4	Collision avoidance
1	Computational complexity
3	Computational methods
1	Computer architecture
2	Computer simulation
1	Computer software
1	Constraint theory
3	Control equipment
1	Control system synthesis
1	Control systems
1	Control theory
1	Data processing
1	Degrees of freedom (mechanics)
1	Design
1	Distributed parameter control systems
1	Electronic guidance systems
2	Error correction
2	Feature extraction
1	Feedback control
4	Flight dynamics
1	Fuzzy control
1	Fuzzy sets
2	Global positioning system
2	Helicopters
1	Hierarchical systems
1	Highway traffic control
1	Identification (control systems)
3	Image analysis
3	Image processing
2	Image segmentation
1	Image sensors
1	Infrared imaging
1	Intelligent control systems

For Help, press F1

Descriptors that both UC Riverside and Penn State share are highlighted in green. Red descriptors co-occur with the selected row item, UC Riverside, but not with the selected column item. Similarly, yellow descriptors co-occur with the selected column item, Penn State, by not with the selected column item.

Here is a complete list of each:

Overlapping terms	UC Riverside terms:	
Air navigation	Approximation theory	ROBOTS_Intelligent
Aircraft	Calibration	ROBOTS_Vision Systems
Artificial intelligence	Cameras	Space applications
Collision avoidance	Computational complexity	Spacecraft
Computational methods	Computer simulation	Speed control
Control systems	Computer software	Standards
Feature extraction	Constraint theory	Stereo vision
Flight dynamics	Control equipment	Underwater equipment
Fuzzy sets	Data processing	Vehicles
		VEHICLES_Navigation
Helicopters	Degrees of freedom (mechanics)	Systems
Image analysis	Design	
Image processing	Error correction	
Mathematical models	Feedback control	
Motion control	Fuzzy control	
Object recognition	Global positioning system	
Obstacle detectors	Hierarchical systems	
Optical flows	Highway traffic control	
Parameter estimation	Identification (control systems)	
Robustness (control systems)	Image segmentation	
Sensor data fusion	Image sensors	
Sensors	Infrared imaging	
Tracking (position)	Intelligent vehicle highway systems	
Video cameras	Kalman filtering	
	Knowledge based systems	
	Manipulators	
Penn State terms:	Motion planning	
Computer architecture	Navigation systems	
Control system synthesis	NAVIGATION_Inertial Systems	
Control theory	Nonlinear control systems	
Distributed parameter control systems	Off road vehicles	
Electronic guidance systems	Optimization	
Maneuverability	Pattern recognition systems	
Military applications	Proximity sensors	
Pattern recognition	Real time systems	
Position control	Remote control	
Process control	Robot learning	
Signal filtering and prediction	Robot programming	
Three dimensional	Robots	

It is interesting to note that many of the shared descriptors have a high expectancy value, indicating that the two institutions may share some ‘niche’ research areas. Here is the same detail window sorted by expectancy:

The screenshot shows the VantagePoint software interface. The main window displays a matrix of expectancy values between various institutions and descriptors. The institutions listed on the left include: Univ of California, Riverside, CA, USA; Texas Transp Inst, Texas A&M Univ, Texas; The Pennsylvania State Univ, University Park; Santa Clara Univ, Santa Clara, CA, USA; Jet Propulsion Lab, Pasadena, CA, USA; Oak Ridge Natl Lab, TN, USA; Universitat der Bundeswehr, Neubiberg; Martin Marietta Astronautics Group, Denver; Lockheed Missiles & Space Co, Sunnyvale; Univ of Technology of Compiegne, Compiegne; Synchrotron Univ, Synchrotron, NY, USA; Office of Naval Studies et de Recherche; Defence Research Agency, Engl; Univ of Southampton, Highfield, Engl; Osaka Univ, Osaka, Jpn; Hitachi Motor Corp, Jpn; Navy Center for Applied Research in Artifi; U.S. Bureau of Mines, Pittsburgh, PA, USA; SONATECH Inc; DLR, Ger; Univ of Girona, Catalonia, Spain; and Univ of Granada, Granada, Spain. The descriptors listed on the right include: Flight dynamics; Obstacle detectors; Object recognition; Optical flows; Computational methods; Aircraft; Calibration; Off road vehicles; Error correction; Helicopters; Tracking (position); Image segmentation; Sensor data fusion; Control equipment; Parameter estimation; Video cameras; Air navigation; Mathematical models; Artificial intelligence; Underwater equipment; Fuzzy sets; Robots; Collision avoidance; Sensors; Image analysis; Image processing; Motion control; Navigation systems; Robustness (control systems); Computer simulation; Control systems; Feature extraction; Global positioning system; Intelligent vehicle highway systems; Motion planning; Real time systems; Vehicles; and Approximation theory. The matrix shows expectancy values for each institution-descriptor pair, with some cells highlighted in green and others in pink.

Using the tools developed for link analysis, the analyst is able to identify and explore a potentially interesting relationship between two institutions that do not appear to collaborate in the area of interest.

Note: Please refer to the following document(s) for full descriptions of these capabilities:

- SPS Document (INDIRECT LINK MATRIX)
- SPS Document (MATRIX LIST)

Conclusions

The goal of this project was to provide a suite of text- and data-mining tools that will help analysts find connections between requirements documents and open-source research literature. Achieving this goal required addressing a range of technical issues including parsing ill-structured documents into structured indexed datasets, data visualization, and advanced analytical techniques.

A model of an analytical process was described in the proposal, and the model serves as a framework within which to structure components of the tools suite and to summarize the conclusions.

Segment the Requirements Document – Multiple approaches were developed to divide the raw dataset into analytical “chunks.” Regular Expression pattern matching proved to be the most powerful and flexible of the automatic approaches. Third-party tools (commercially available) offer some capabilities, especially useful for PDF formats,

although they require some user intervention. User-defined segments have been introduced as a backup approach for use when patterns cannot be reliably defined. “Content measures” (e.g., change in discourse) were less successful in this program, although tools have been built into TechOASIS that may be used to research these approaches.

Extract Metadata – The capability to extract metadata was greatly enhanced by the introduction of the Regular Expression based Import Engine. Two other approaches added unique functions. “Import Variables” enable metadata within the scope of the source document but outside the scope of the segments to be associated with the segments. “FrankenRecords” combine metadata from diverse, external data sources to be associated with segments.

Parse the Segments – As with other functions, this function also leverages the integration of Regular Expression pattern matching into the Import Engine. The suite of 15 commands may be combined in many different configurations to provide a wide array of parsing and transformational steps during the import process. An additional approach evolved late in the project – using the natural language parser to parse noun phrases that occur in proximity (as defined by a sentence range) to a set of terms of interest. While not a part of the “import” process, this analytical tool results in a contextually rich set of phrases that are closely related to the terms of interest.

Extract Entities – Automatic and user-confirmed entity extraction is built around user-defined dictionaries (normal text or Regular Expression). Users report that one of the most powerful components of the tool suite developed under this program is the ability to “protect” entities during the NLP parsing process, which enhances frequency counts for context-rich phrases. Additionally, Boolean, proximity, and order-dependent searching of free text fields have proved to be beneficial for identifying and tagging (grouping) ill-structured records.

Reduce and Combine Data – Data normalization (cleanup) and transformation (thesaurus) make a big difference in the quality of analysis. However, automatic processes are of limited effectiveness because of the great variation in the raw data. Of greatest benefit in this stage of the analysis are tools that augment the analyst’s task – performing and managing the mundane activities, while making it as convenient as possible for the analyst to add value to the data. Numerous capabilities have been added, but among the most significant for many users is the ability to (a) stop and resume the cleanup tasks and save, (b) save and merge cleanup operations as thesauri, and (c) normalize/reduce thesauri.

Decompose Segments – Two primary tools advance this activity, and both of them are based on cross-field analysis. Graphical pop-ups of categorical data (and other types of data) on maps provide the ability to compare and contrast the segments that fall into different clusters. Similarly, “expectancy arrows” in Detail Windows highlight out-of-the-ordinary results.

Mine the Source Document – The process of iteratively mining the source document is enhanced by enabling the primary analytical processes (e.g., mapping and PCD) to occur on subsets of records while operating within the full, primary dataset. Other clustering capabilities were explored, and third party tools (e.g., CLUTO, from the University of Minnesota) have been built in for research purposes. Finally, Detail Windows and export of 3-field pivot tables enable three-field visualization in support of this mining activity.

Refine the Query – A key aspect of query refinement is the selection of the feature space (or vector space) within which to rate query terms. We chose a commonly used technique known as Term Frequency Inverse Document Frequency (TFIDF) to help the user select these terms.

Mine the Open Literature Dataset – The process of relating the language in the requirements document to the language in the open research literature led to exploring ways to find synonyms in the two term sets. Our eventual approach is built on Princeton's WordNet thesaurus; however, the size of the thesaurus presents significant practical limitations to its usefulness in this application. Further research is needed in this area.

During the course of this project, two contract modifications added new directions to our work scope. Among them, the following merit particular mention:

Leveraging Open-Source Repositories and WebQL Interface – The much-publicized “freely accessible” data sources available on the Internet offer a tempting target for these types of analysis. Fee-based services can be quite expensive, and significant cost savings can be realized if equivalent data can be freely obtained. Some of the added tasks sought to build bridges to these free data sources. It was much more difficult than it seemed it should be. The conclusion of the matter is that many (if not most) of these “open” data providers currently prohibit (albeit only administratively) this type of access/crawling/mining of their data. A later adjunct to this effort added a bridge to a commercially available tool for accumulating data from Internet-based sources – WebQL. With this bridge, a TechOASIS user can (a) access Internet data sources for which a query has been pre-defined and (b) automatically import the data into TechOASIS.

Quick XML Import – Enabling import of XML data was accomplished very early in the project. However, later tasking created a two-step process for importing XML data from virtually any data source. This holds great promise, as XML becomes a pervasive format for data delivery.

Simultaneous Cross Dataset Analysis – One approach to analyzing multiple datasets as a collection is to merge the datasets into one large dataset. This becomes inefficient if the datasets are diverse – perhaps sharing only one field in common. The ability to simultaneously analyze multiple diverse datasets without fusing them offers significant improvements in efficiency.

Link Analysis – Among the last and most promising activities of this project is link analysis. The goal is to improve the chances of finding “hidden” links in the data – for example finding a link between “A” and “C” when they are never directly related to each other, but both are related to “B”. The Indirect Link Matrix and associated tools form a powerful tool for exploring and discovering this type of relationship.

Recommendations

TechOASIS has not yet been widely adopted by DoD technology managers. The utility of TechOASIS is indicated by the success of VantagePoint in the commercial market. Worldwide, dozens of companies use VantagePoint and Thomson’s Derwent Analytics – the user base numbers in the hundreds. At the close of this project, there is emerging interest in TechOASIS in the government sector outside of DoD.

What can be done to realize the benefit of TechOASIS within DoD, and the Army in particular?

- Integrate into the culture. We have repeatedly encountered resistance to TechOASIS due to two factors. First, the “expert” factor. This is typified by quick dismissal of results from TechOASIS as “obvious” to an expert in the field. Second is the “relationship” factor. This is characteristic of those that gain their intelligence by going to conferences and talking with other experts, and view a tool like TechOASIS as taking that away. Both are valid critiques. TechOASIS should be understood and presented as a component in a broader process that includes both experts and relationships among experts. It should be shaped as a component of the culture, augmenting expertise and conferences.
- Develop a framework that captures the issues and questions. Understanding the issues and questions facing DoD technology managers is essential to realizing broader benefit from TechOASIS. A parallel might be drawn with developing useful information products for technology managers in commercial enterprise (see Tech Mining, Porter and Cunningham, 2005). Porter and Cunningham propose a framework of 14 Issues → 39 Questions → 150 Indicators → “Rapid Technology Information Products,” which are one-page displays of the indicators surrounding a particular issue or question. Their framework comes from commercial enterprise, but some directly apply to DoD and many others may be straightforwardly tailored.
- Simplify and specialize TechOASIS. The end result of this project is a powerful, flexible suite of tools. The tools are well integrated at the data level, and the user-interface level is generally internally consistent. However, the complexity can be overwhelming to a beginner. TechOASIS contains a sufficiently large suite of tools that it may be beneficial to tailor it to several “vertical” applications, each possessing a subset of the full suite but specializing in a particular domain, issue, or question. Developing the framework for these applications is an essential first step (see the prior point).

- Establish an infrastructure to support technology managers. For technology managers to benefit from these tools and data, they need at least four things: (1) Access to technical data. Commercial and DoD sources for technical data abound. However, the access seems to be fragmented at best. (2) Tools. Because TechOASIS was developed under the SBIR/STTR program, it is available at no cost to government users for government purposes. The Army sponsor has the full product that may be distributed under these terms. (3) Training. During this project, the Army has funded Search Technology to provide periodic TechOASIS training. Search Technology is taking steps to establish a TechOASIS training course on a GSA schedule. (4) Support. During this project, the Army has funded Search Technology to provide technical support to government users. Going forward, Search Technology will provide technical support on an annual subscription basis (again on a GSA schedule). The pricing and terms will be the same as those offered to VantagePoint commercial customers. This subscription includes web-based distribution, telephone and e-mail technical support, maintenance, user support (import filters), and software updates and upgrades as they become available.

References

Frey, P., “TechOASIS – Software Product Specification,” Contract DAAE07-02-C-L016: An Integrated Suite of Text and Data Mining Tools – Phase II. Search Technology, Inc., August 2005.

Frey, P., Porter, A., Newman, N., and Minsk, B. “Technology Opportunities Analysis System (TOAS): Phase 3 - Final Technical Report.” Contract DAAH01-98-C-R128 Technical Report, Search Technology, Inc., July 2001.

Li-Ping Jing, Hou-Kuan Huang, Hong-Bo Shi, “Improved Feature Selection Approach TFIDF in Text Mining,” IEEE, Proceedings of the First International Conference on Machine Learning and Cybernetics, Beijing, Nov 2002.

Porter, A.L., and Cunningham, S.W., Tech Mining: Exploiting New Technologies for Competitive Advantage, Wiley, New York, 2005.

“VantagePoint Users Guide – version 4.1,” Search Technology, Inc., August 2005.

“VantagePoint On-Line Help – version 4.1,” Search Technology, Inc., August 2005.

“WordNet, a lexical database for the English language,” <http://wordnet.princeton.edu>, Cognitive Science Laboratory, Princeton University. August 2005.

Appendix A

Mining Conference Proceedings for Corporate Technology Knowledge Management

Robert J. Watts¹ and Alan L. Porter²

¹U.S. Army Tank and Automotive Research, Development & Engineering Center, USA

²Search Technology, Inc., Norcross, Georgia, USA

Abstract—An organization's knowledge gained through technical conference attendance is generally isolated to the individual(s) attending the event. The aggregate corporate knowledge is extremely limited, unless the organization institutes a process to document and transfer that knowledge to the organization. Even if such a process exists, the knowledge gains are limited to the experiences and communication skills of the individuals attending the conference. Many conference proceedings are now published and provided to attendees in electronic format, such as on CD-ROM and/or published on the internet, such as IEEE conference proceedings listed at http://www.computer.org/proceedings/proceed_a-h.htm.

These proceedings provide a rich repository that can be mined. Paper abstract compilations reflect "hot topics," as defined by the researchers in the field, and delineate the technical approaches being applied. R&D profiling can more fully exploit recorded conference proceedings' research to enhance corporate knowledge. This paper illustrates the potential in profiling conference proceedings through use of WebQL information retrieval and TechOasis (VantagePoint) text mining software. It shows how tracking research patterns and changes over a sequence of conferences can illuminate R&D trends, map dominant issues, and spotlight key research organizations.

I. INTRODUCTION

How does one keep up with R&D? Information is spewing forth at ungodly rates. Multiple access modes bring this information to your fingertips, spilling over your desk, into your coffee cup, and threatening to drown us all. But, rescue is at hand – new tools enable powerful analyses and information visualizations [1, 2, 3]. Furthermore, these can be directly pointed to answer your pressing management of technology (MOT) questions [4, 5, 6, 7, 8, 9].

One key venue for exchange of fast-breaking research developments is "the conference." Just as PICMET exposes you to the latest explorations in MOT, manifold conferences address all sorts of technical issues. In this paper we illustrate how to gain value-added information from conferences. We explore alternative data access modes and what these can offer technology managers. In that there is no free lunch in the data world, we want to compare what it takes to obtain useful MOT intelligence from a) free web versions of the data vs. b) obtaining the proceedings abstract records via pay databases such as EI Compendex and INSPEC.

II. TECHNICAL APPROACH

Having access to the IEEE website, and professional interest in certain of its topics, we began our investigation there. We searched the IEEE list of conference proceedings for specific topics (e.g., noun phrases within the full conference name) to locate those covering topics of special interest. For this application, we selected a particular conference – IEEE International Workshops on Database and Expert Systems Applications (DESA). We are interested in their coverage from 2001-04 (four conferences).

We used WebQL [<http://www ql2.com/>] to mine the IEEE Conference Proceedings web site [http://www.computer.org/proceedings/proceed_a-h.htm]. WebQL, from QL2 Software, is a software tool enabling quick development and easy deployment of software agents to extract data from the World Wide Web and many other unstructured data sources. We thus identified the conference proceedings of interest and corresponding web links to be mined (using a second WebQL script). We focused our conference listing search on expert systems and discovered the IEEE conferences, "International Workshops on Database and Expert Systems Applications (DESA)." The second WebQL script searched for and retrieved specific conference proceedings web link information and compiled it in Excel format. Figure 1 presents a screen capture of the IEEE web site page that provided the links mined to retrieve the IEEE abstracts analyzed here.

Once extracted, WebQL can structure the data into standard output file formats – HTML, PDF, DOC, etc. We formatted the retrieved conference listings for Excel file analysis. This WebQL output file could have simply been viewed and searched in Excel. However, we imported the file into Tech OASIS¹ to facilitate richer analyses. Each paper's summary information included: conference name, conference date, conference location, paper title, authors name(s), author(s') affiliation(s) and the paper abstract.

The Tech OASIS Excel quick import engine/filter was edited to provide natural language processed (NLP) text fields for both the paper title and abstract fields. Use of NLP-parsed terms and phrases provides a way to mine the actual content of the abstracts. Through NLP text profiling we can get at the topics researchers are pursuing. Our targets include knowledge about the entire research domain of interest, including:

¹ Tech OASIS is for U.S. Government use. The commercial versions of this software are *VantagePoint* [www.theVantagePoint.com] and *Derwent Analytics* [<http://www.derwent.com/products/dapt/derwentanalytics/>].

* *what* – what are the hot topics?
 * *who* – who are the research leaders on particular topics?

* *where* – where are the centers of knowledge?
 * *when* – what are the trends in research?

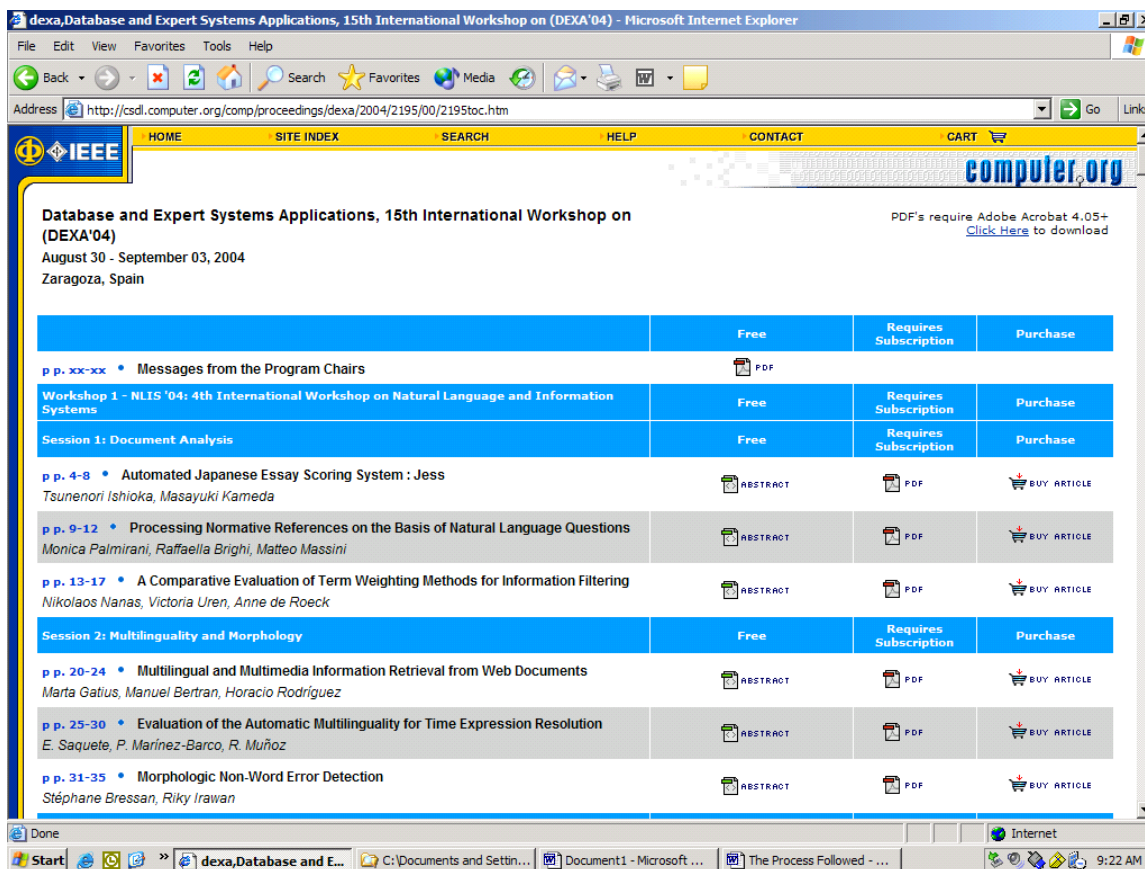


Fig. 1. IEEE web site mined by WebQL software.

NLP algorithms capture useful chunks of text within the free-text portion of the abstracts. We have found that certain text-processing-aids greatly improve the quality of the information available. By borrowing keywords from indexed databases we assure that domain-specific terms and phrases are captured in the free text. For instance, if we are interested in “expert systems,” we don’t want the NLP parser to separate the terms into “expert” and “systems.”

To identify informative terms in our analyses, we used a 3-step process. First we examined a limited set of abstracts containing research domain terms and phrases within the conference proceedings. Domain-specific terms and phrases from that source formed the basis for a search strategy for a second source -- indexed databases (EI Compendex and INSPEC). The descriptors and identifiers (i.e. the index terms or keywords) from the indexed databases were compiled to create an improved set of terms and phrases for the domain under study. In the third step, these terms and phrases were tagged in the conference proceedings’

abstracts files and extracted (i.e., protected during NLP processing on import into Tech OASIS). This resulted in a contextually rich set of entities on which to profile the conference proceedings. Put another way, we “borrowed” the index terms from EI Compendex and INSPEC to help analyze the version of the conference proceedings gathered directly from the website (that lacks index terms).

We began this process by looking for clues. The 2001-04 International Workshops on Database and Expert Systems Applications cover many topics, so devising a suitable search strategy to retrieve corresponding information from huge databases was not trivial. Our “Rosetta Stone” appeared in the 2002 DESA proceedings in the form of a sequence of messages from session chairpersons. These consisted of descriptive abstracts by the co-chairs about the sub-workshops on: holonic and multi-agent systems (HoloMAS), electronic business hubs (WEBH), trust and privacy in digital business (TrustBus), negotiations in electronic markets (e-Neg), mobility in databases and

distributed systems (MDDS), theory and applications of knowledge management (TAKMA), management of information on the web (MIW), web based collaboration (WBC), natural language and information systems (NLIS), web semantics (WebS), and very large data warehouses (VLDWH). The text from these co-chairs'

messages was manually scanned and the terms and phrases in the search strategy, Table 1, were identified. This search strategy uses Boolean logic to search EI Compendex and INSPEC. Closed parentheses mean that the terms are required to be adjoining. The question mark indicates wild card character(s).

TABLE 1. DATABASES AND EXPERT SYSTEMS SEARCH STRATEGY

Set	Items	Description
S1	1297161	PY>2000
S2	3150	S1 AND (EXPERT()SYSTEM?)
S3	5566	S1 AND (MULTI-AGENT?)
S4	3073	S1 AND (DISTRIBUTED()SYSTEM?)
S5	48	S1 AND ((ELECTRONIC()MARKET?) AND NEGOTIATION?)
S6	1519	S1 AND ((COLLABORATIVE OR GRID)()COMPUTING)
S7	2927	S1 AND (KNOWLEDGE()MANAGEMENT)
S8	4391	S1 AND (SOFTWARE()AGENT?)
S9	501	S1 AND (TEXT()MINING OR SUMMARIZATION OR CATEGORIZATION))
S10	19216	S2 OR S3 OR S4 OR S5 OR S6 OR S7 OR S8 OR S9
S11	41752	S1 AND (INTERNET OR WWW OR (WORLD()WIDE()WEB))
S12	1372	S1 AND (WEB(2N)INFORMATION)
S13	278	S1 AND (WEB(2N)COLLABORAT?)
S14	967	S1 AND (WEB(2N)SEMANTIC?)
S15	42743	S11 OR S12 OR S13 OR S14
S16	3067	S10 AND S15

Had we not found this set of messaging telling about the workshop themes, we would have considered two other ways to generate search terms to use in the databases. One approach is to list the NLP title phrases and highlight defining terms therein. Another is to locate abstracts within the workshop whose titles and/or texts suggest over-viewing – e.g., “forecast of,” “technology assessment,” “new trends,” and so forth.

Our second step applied the Table 1 search strategy to retrieve 3067 and 1344 abstracts, respectively, from the INSPEC and EI Compendex databases (as licensed from Dialog, Inc., a database provider). A combined list of descriptors and identifiers was compiled from the two Dialog search files.

For our third step, this list was used to extract domain-specific terms via another import of the web-sourced IEEE conference proceedings. Compared to files compiled using the standard Tech OASIS import engine, the resulting abstract files had more than triple the number of abstract NLP terms and phrases available for cluster analyses. This demonstrates the utility of

applying index terms (keywords) from outside sources. It also shows the value in protecting those terms during natural language parsing. The results were

- The 2001 proceeding abstracts' extracted NLP lists had 335 terms with record frequencies greater than 2 (208 were descriptor/identifier domain specific entities) vs. 91 terms compiled by the standard NLP processed list.
- The 2002 proceeding abstracts' had 454 such terms (263 entities) vs. 114 for the standard NLP import
- The 2003 proceedings had 316 (195 entities) vs. 81 and
- The 2004 proceedings had 336 (207 entities) vs. 102.

We next describe how these enriched terms were used to profile the IEEE DESA Proceedings. This results in qualitatively richer understanding of the content of these conferences. It enables users to understand overall research emphases, as well as to pinpoint papers of particular interest.

III. ABOUT PROCEEDINGS' ABSTRACTS

The WebQL web crawler software retrieved 148, 152, 157 and 173 abstracts, respectively, for the 2001, 2002, 2003 and 2004 IEEE Databases and Expert Systems Application (DESA) conference proceedings. We analyzed the four annual proceedings separately and combined. Managers can gain insights on research "hot topics" by analyzing the individual proceedings. The combined proceedings file provides information on topical trends and regular attendees.

Table 2 shows the leading organizations vs. conference dates. Such a compilation provides knowledge about who regularly presents at these conferences. This can point us toward cutting edge researchers. For instance, were we planning to send someone to attend the next DESA workshop, we might expressly point them to make contact with the Czech Technical University and University of Greenwich researchers. Observing Table 2, one also observes that

foreign sources dominate publication of research at this forum. Is this observation true for the broader field of research?

Table 3 shows the topical emphases of the leading organizations at DESA. These reflect term clusters (or factors – we apply principal components analysis to the NLP extracted entities, terms and phrases). This provides information on research focus areas of each organization. The leading conference presenter (Czech Technical University) concentrates on three primary areas: heritage, interoperability and multi-agents. The Open University's abstracts primarily cluster in only one area -- the heritage factor. Five of six of Hewlett-Packard's abstracts fall in the business factor group. This intelligence would support decisions on who might make attractive collaborative partners. Interesting observation -- four factor groups (authentication, evolution, electronic commerce and e-government) have only one lead organization with more than one abstract. So, the technology manager seeking expertise at this venue has clear targets.

TABLE 2. LEADING AFFILIATIONS AT IEEE DATABASES AND EXPERT SYSTEMS APPLICATIONS

		# Records	173	157	152	148
# Records	Affiliations (Cleaned 2)	2004	2003	2002	2001	
15	Czech Technical University, Prague, Czech Republic	3	4	6	2	
12	University of Greenwich, London, UK	8	4			
9	University of Vienna, Austria	3	3	2	1	
8	Vienna University of Technology		2	1	5	
8	Open University, Milton Keynes, UK	3	2	3		
8	Poznan University	1	3	2	2	
7	Nanyang Technological University		2	2	3	
6	Tohoku University, Japan		2	2	2	
6	University of Linz, Austria	1	1	1	3	
6	University of Pittsburgh, PA	2	1	2	1	
6	Hewlett-Packard Corporation	2		2	2	
6	Tokyo Denki University, Japan	2	2	1	1	
6	Fraunhofer AIS, Germany	3		3		
5	Middlesex University	1		1	3	
5	Università di Milano, Italy	3		2		
5	Yamagata University, Japan	2		1	2	
5	National Technical University of Athens			1	4	
5	University of Calgary			3	2	
5	Toyo University, Japan	2	1	1	1	
5	University of Zaragoza, Spain	3			2	
5	Monash University		1	2	2	
5	Iwate Prefectural University	2	1		2	
5	Johannes Kepler University Linz, Austria	2	1	1	1	
4	Fukuoka Institute of Technology (FIT), Japan	3	1			
4	University of Alberta, Edmonton, Canada	1	1	1	1	
4	University of Oklahoma		2	1	1	
4	University of Missouri-Kansas City, Kansas City	1	2		1	
4	Università di Brescia, Italy	1	1	2		
4	Imperial College London	3	1			
4	University of Tokyo			2	2	
4	City University of Hong Kong		2	1	1	
4	University of Montreal			2	2	
4	Univ. de Castilla la Mancha, Spain	4				
4	University of Malaga, Spain	2	2			

TABLE 3. LEADING AFFILIATIONS VS. FACTOR MAP GROUPS

		ABSTRACT (NLP) C:Entities (factors)																	
	# Records	136	118	113	105	83	83	68	64	59	51	43	42	41	36	32	28	26	20
# Records	Affiliations (Cleaned 2)	datasets	retrieval	interoperability	traffic	query	video	Business	authentication	multi-agent	learning	distributed	mobile	knowledge management	evolution	real-time	electronic commerce	Heritage	e-government
15	Czech Technical University, Prague, Czech Republic		3	5			2			4				3	2			6	
12	University of Greenwich, London, UK	3		2			2	2				2		2					
9	University of Vienna, Austria	3	3		3			2			2	4		2					
8	Vienna University of Technology	5	2	2		2													
8	Open University, Milton Keynes, UK		2				2											6	
8	Poznan University	3		2	3	3		2											
7	Nanyang Technological University		2				3										2		
6	Tohoku University, Japan	2			3			2											
6	University of Linz, Austria		2																
6	University of Pittsburgh, PA													3					
6	Hewlett-Packard Corporation							5											
6	Tokyo Denki University, Japan				3														
6	Fraunhofer AIS, Germany				3														
5	Middlesex University	2	2								3								
5	Università di Milano, Italy	2		2	3														
5	Yamagata University, Japan				4														
5	National Technical University of Athens						2	3						2					2
5	University of Calgary				2					5						3			
5	Toyo University, Japan		4																
5	University of Zaragoza, Spain												2						
5	Monash University												3						
5	Iwate Prefectural University		2		2		2												
5	Johannes Kepler University Linz, Austria											2							
4	Fukuoka Institute of Technology (FIT), Japan		3		3														
4	University of Alberta, Edmonton, Canada	3																	
4	University of Oklahoma				2								2			2			
4	University of Missouri-Kansas City, Kansas City																		
4	Università di Brescia, Italy	2		3															
4	Imperial College London				2														
4	University of Tokyo						2												
4	City University of Hong Kong		2		2						2								
4	University of Montreal																		
4	Univ. de Castilla la Mancha, Spain																		
4	University of Malaga, Spain								2										

IV. FACTOR MAP CLUSTER GROUPS

Although the messages from the co-chairs of the 2002 conference were most useful in developing our database search strategy, they appeared to bias the clustering of the conference proceeding abstracts. Therefore, the co-chair messages were removed from the files before Table 3 was derived. However, Table 4 shows in what groups the co-chair messages clustered during initial analyses. This knowledge helps verify the process for using the NLP-extracted entities to derive the factor groupings.

During the first iteration, eleven factor groups were derived, as shown in the 2nd and 5th columns of Table 4 and preceded by “Map:”. The remaining terms in the 2nd and 5th columns are the other terms of the respective factor group. For example, the factor group, Map: University, consists of all abstracts containing terms: technology, research, messaging, university, topic or exchange. All of the co-chair messages clustered together in this University factor group. Additionally, the University factor group contained 190 of the 630 published abstracts. This large number influenced the decision that the co-chair messages were biasing the factor analysis.

However, observing the clustering of the messages in the other ten factor groups helps validate the NLP entities extraction and standard factor map process used to cluster the abstracts. Viewing Table 4, the “trust” factor, defined by the terms: control, security, privacy, trust, access control and authentication, had the highest loading abstract (i.e., Hi-Load (-0.69)). The co-chair message for the Trust and Privacy in Digital Business working group (TrustBus), the

only co-chair message to be clustered in the trust factor group, had a loading coefficient of -0.61, and thus appears appropriately grouped. Two co-chair messages, Web Based Collaboration (WBC) and Mobility in Databases and Distributed Systems (MDDS) had loading coefficients of -0.34 and -0.25 in the agent factor group, which had a highest abstract loading coefficient of -0.6. A level of confidence in the proceedings analysis process can be gained by comparing the factor defining terms and the high loading co-chair messages shown in Table 4.

TABLE 4. FACTOR MAP GROUP DEFINING TERMS (COMBINED 2001-04 IEEE PROCEEDINGS)

# Records	Map: University	Hi-Load (1.63)	# Records	Map: commerce	Hi-Load (0.54)
101	technology	WBC (1.63); MDDS (1.36)	31	commerce	WEBH(0.4)
92	research	NLIS (1.02); MIW (0.92)	14	electronic commerce	e-Neg(0.34); TrustBus(0.28)
25	messaging	HoloMAS(0.9); WebS(0.64)	# Records	Map: XML	Hi-Load (-0.51)
21	University	TAKMA(0.61); e-Neg(0.59)	49	XML	MDDS(-0.5); WEBH(-0.5)
19	topic	TrustBus(0.56)	38	standard	HoloMAS(-0.46)
14	exchange	VLDWH(0.49); WEBH(0.45)	20	data model	WebS(-0.41)
# Records	Map: trust	Hi-Load (-0.69)	# Records	Map: Heritage	Hi-Load (-0.85)
45	control	TrustBus(-0.61)	64	Ontology	WebS(-0.26)
37	security		17	Heritage	
18	privacy		13	exploration	
18	trust		12	cultural heritage	
14	access control		# Records	Map: datasets	Hi-Load (0.99)
13	authentication		49	analysis	VLDWH(0.58)
# Records	Map: agent	Hi-Load (-0.6)	16	storage	WBC(0.18)
38	mobile	WBC(-0.34)	15	cluster	
33	agent	MDDS(-0.25)	14	grids	
20	server		13	data warehouse	
16	mobile agents		12	data mining	
# Records	Map: multi-agent	Hi-Load (1.79)	12	datasets	
48	agents	HoloMAS(1.79)	# Records	Map: information retrieval	Hi-Load (-0.60)
18	agent system	WBC(0.66)	57	documents	NLIS(-0.43)
18	multi-agent		48	search	
14	manufacturing		37	retrieval	
# Records	Map: traffic	Hi-Load (0.94)	17	information retrieval	
31	algorithms	MDDS(0.94)	# Records	Map: learning	Hi-Load (0.40)
19	real-time	HoloMAS(0.62)	41	learning	MIW(0.30); WebS(0.29)
15	traffic		14	e-learning	NLIS(0.16)

The factor analyses of the NLP extracted entities were redone, excluding the co-chair message abstracts. The factor map of the combined file (2001-04) of proceedings' abstracts is shown in Figure 2. Each factor, represented as a node, has a drop-down box containing

the group-defining terms. When viewed together, these hi-loading terms help provide a better understanding of the concepts documented in the grouped abstracts. Links between nodes show factors that relate more closely to each other.

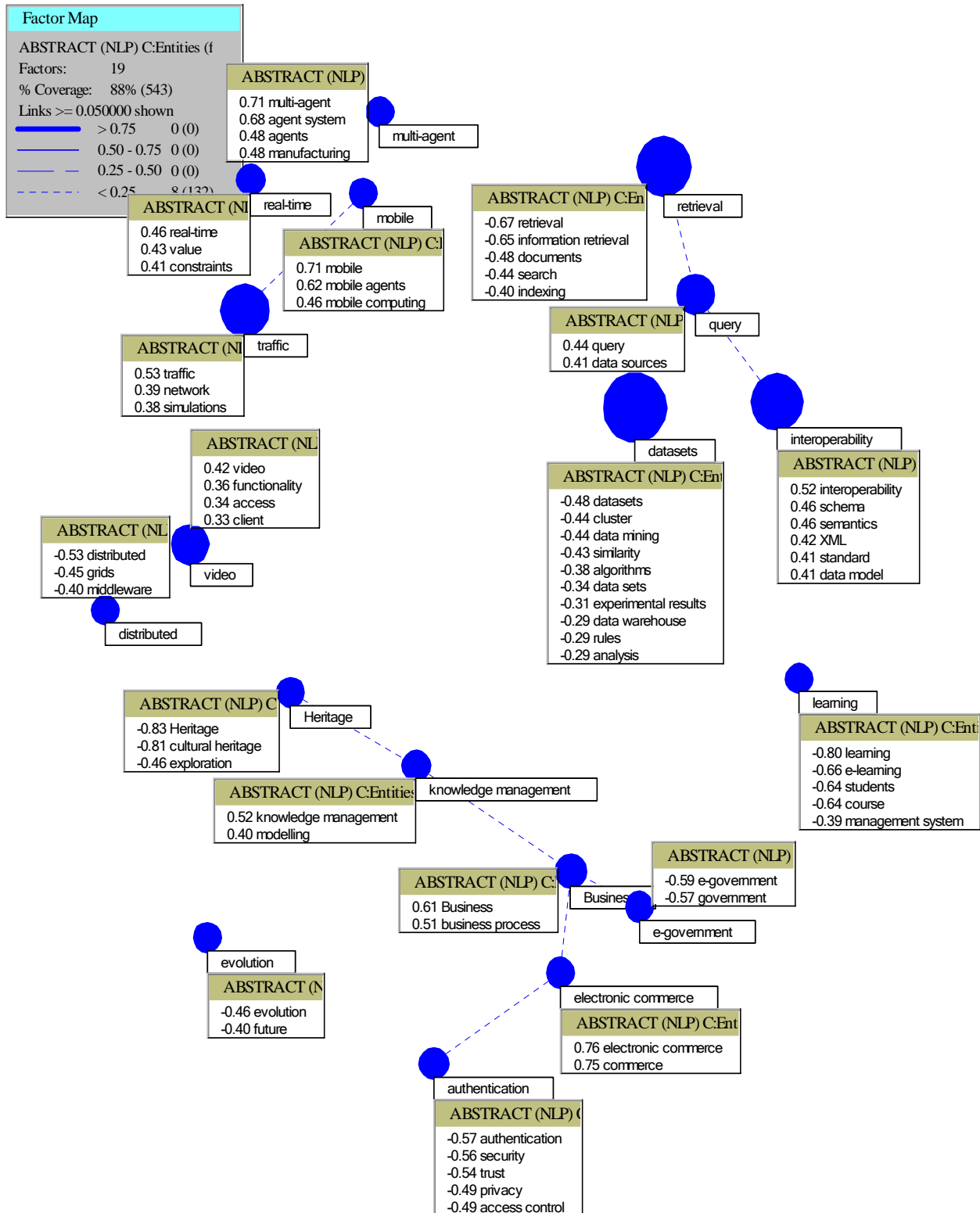


Fig.2. Factor Map of Abstract NLP Entities – IEEE Databases and Expert System Applications Conference Proceedings 2001-2004 (No 2002 Messages).

Figure 3 provides the histograms for each of the Figure 2 factor groups and the number of abstracts presented annually. Such charts can provide managers intelligence on which sub-disciplines dominate the conference subject matter and which categories of research are declining or rising. For example, publications in e-government, electronic commerce and the business

factor groups have declined over the four-year period. Experts in the field could best explain the reasons for the declining research; perhaps, applications have increased (technology matured) and need for research declined. Similarly, one can observe that the five most active areas of research in the 2004 conference were retrieval, interoperability, traffic, query and video.

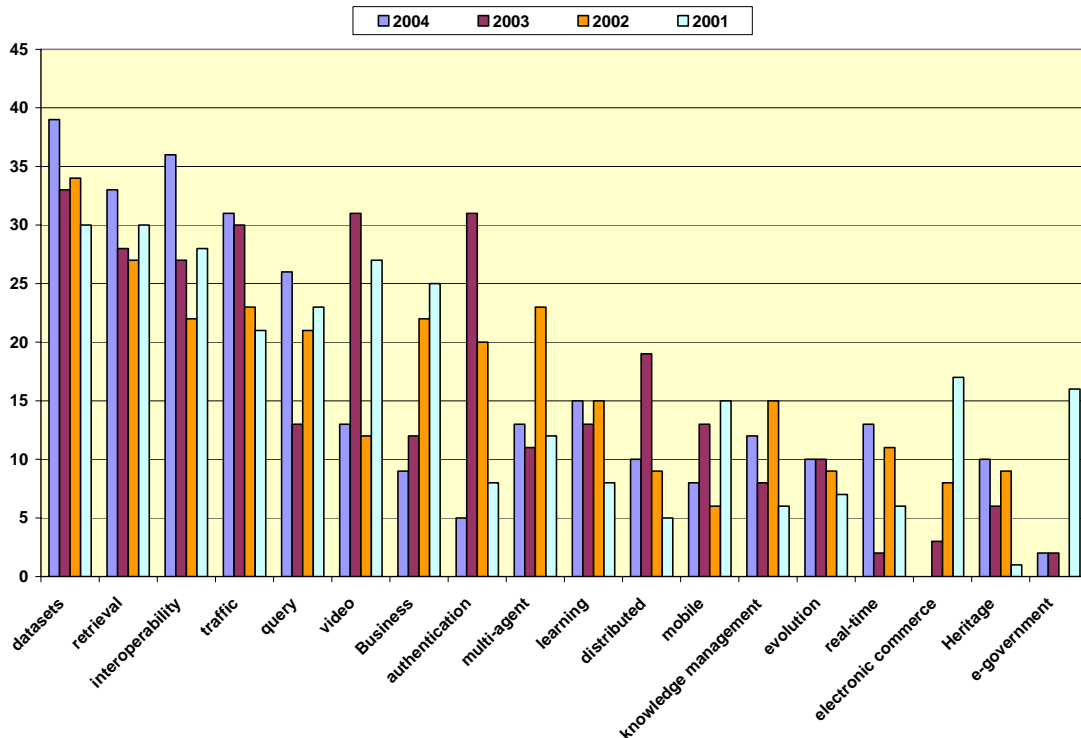


Fig. 3. IEEE Databases and Expert System Applications Factor Groups' chronologies.

Figure 4 depicts the factor map for the 12% outliers – the abstracts that were *not* clustered in the factor mapping (Figure 2 depicts 88% of the abstracts). Factor map groups represent consensus term usage. Abstracts not using these consensus terms may represent new research topics.

Let's explore Figure 4. The term "autonomic computing" appears in two factor groups. Autonomic computing occurred first in 2003 in 7 abstracts and then in 4 abstracts in 2004. In a 2003 paper, Constantinescu states "Systems which are autonomic, capable of managing themselves are required" in "Towards an Autonomic Distributed Computing System." In a 2003 paper, Sterritt et al. claim autonomic computing aims to (i) increase reliability by designing systems to be self-protecting and self-healing; and (ii) increase autonomy and performance by enabling systems to adapt to changing circumstances, using self-configuring and self-optimizing mechanisms. This field, autonomic

computing, appears to fit the definition of an emerging area of research.

By mining down to individual abstracts that have been self-organized into topical groups, managers can quickly gain insights on the "hot topics." Through such mining in Autonomic Computing, we find that an application needs to be aware of its environment. In the 2004 paper, "Simulation Model for Self-Adaptive Applications in Pervasive Computing," Huebscher et al. state "While the term "environment" is not normally understood as being a physical environment, in Pervasive Computing many applications do actually need to monitor the physical environment in which they are deployed." The profiled conference proceedings can, thus, provide both a "meta-perspective" – a bird's eye view (e.g., who are the leading publishers, what are the central research focus areas, etc.), and targeted access to specific information.

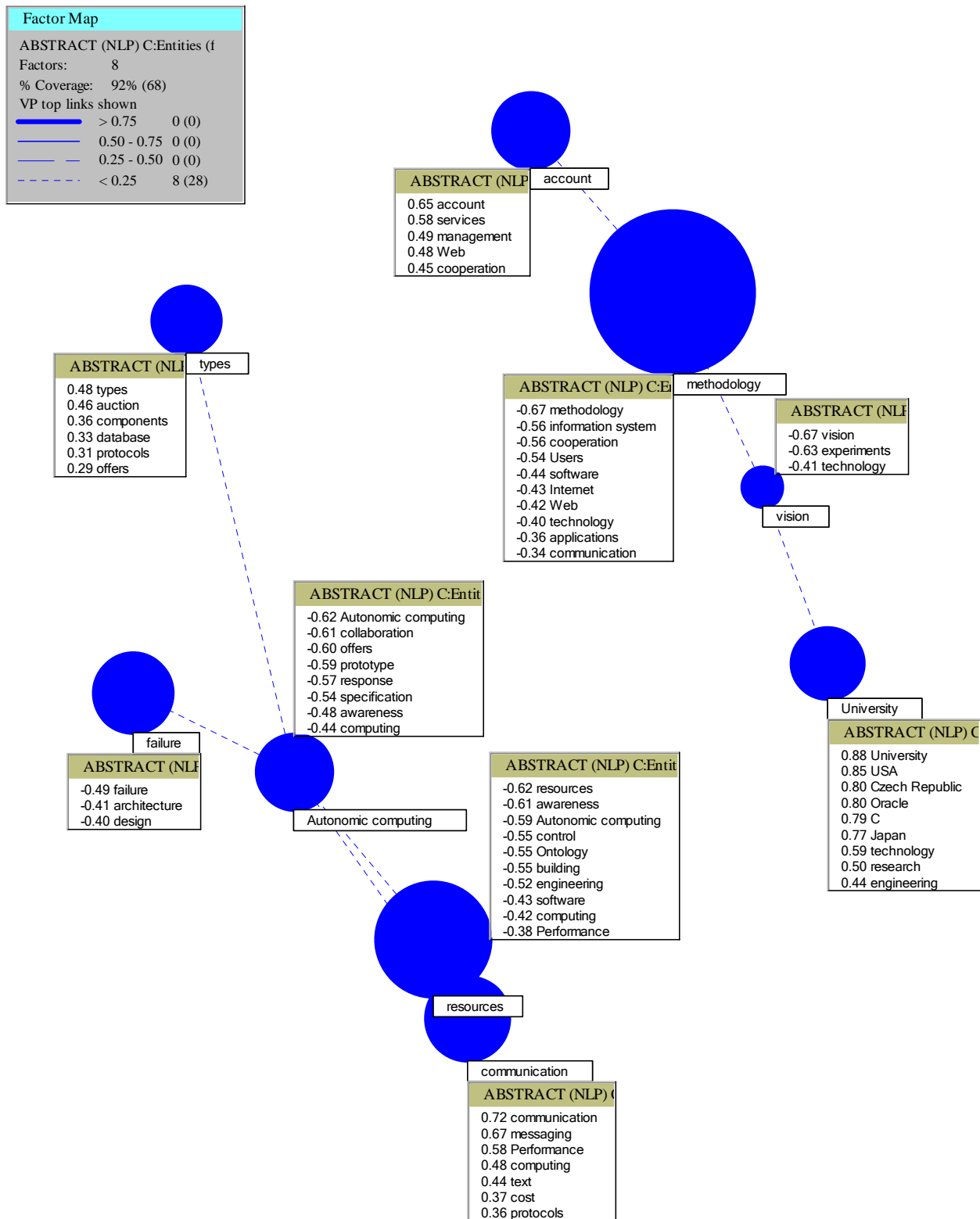


Fig.4. Factor Map of NLP Entities for Non-factored Abstracts – IEEE Databases and Expert System Applications Conference Proceedings 2001-2004 (No 2002 Messages).

V. CONCLUSIONS

We both demonstrate and begin verifying a process to profile non-indexed free-text information (i.e., web-accessible conference proceedings abstracts). We tag and protect lists of research domain-specific terms, compiled from indexed databases (e.g., EI Compendex, INSPEC), within the abstracts' free text. The tagged entities are extracted during NLP parsing of the abstracts to compile a contextually rich set of terms and phrases on which to profile the free-text documents. To accomplish this process, we briefly introduce and use WebQL information retrieval and TechOasis (VantagePoint) text mining software.

More importantly, we demonstrate how the profiled conference proceedings can be used by technology managers. Specifically, intelligence about the conference research domain can be derived, including:

- * *what* – what are the hot topics?
- * *who* – who are the research leaders on particular topics?
- * *where* – where are the centers of knowledge?
- * *when* – what are the trends in research?

We suggest analytical approaches to further validate the demonstrated analysis approach. The standard PCA factor analysis process uses a metric comprised of the population percentage clustered and cluster quality measures (entropy, F-measure and cohesiveness) [10]. Further research should compare cluster quality measures for the factor groups of alternative approaches.

Cluster analysis strives to create "highly internally homogenous groups, the members of which are similar to one another, and highly externally heterogeneous groups, members of which are dissimilar to those of other groups" [1]. Steinbach et al. [2] discuss and apply measures of cluster quality, both internal and external measures of "goodness."

For this example, we observed lower entropy and F-measures for the factor groups derived from the indexed database for the 2004 conference proceedings than obtained for the NLP entity extracted factor groups. This implies that analysis of indexed data provides better factor groups; but that indexing using external information takes time and resources. However, the NLP entity extraction process clustered the same percentage, 97%, of the 2004 abstracts into factor groups. In contrast, the factor groups, created by the standard NLP abstract terms analysis approach, clustered only 66% of the 2004 abstracts. In addition, the 2004 proceedings abstracts' yielded 336 terms with record frequencies greater than 2 (208 were descriptor/identifier domain specific entities) vs. 102 terms compiled by the standard NLP processed list and 149 available for the indexed terms database, EI Village. The NLP entity extraction process, in this case

study, provided the greater number of terms for the factor group analysis. One could argue that it is difficult to make all-inclusive assignments of indexed terms for the abstracts and having self-assignment through entity extraction provides the more thorough approach. Further research should assess this claim.

We note an advantage of using WebQL to retrieve the information to be analyzed. Using WebQL, we could tailor the information, both content and format, to meet our analysis needs. Licensed database suppliers, on the other hand, must provide a set of standard data formats to meet the majority of customer information processing needs. The tailored retrieved information required less cleaning and provided more on-target field lists summaries.

We note the IEEE Databases and Expert System Applications proceedings contained mostly foreign-sourced research and wondered whether this was true for the broader field. This question begs further research. In a complimentary and more general vane, research on how to gauge conferences as to how well, statistically, they reflect the broader field of research might be of value to technology managers.

Finally, information profiling can support other technology management issues to allow a manager to:

- Assess another organization's strengths and weaknesses (e.g., to refine decisions on merger and acquisition)
- Assess one's own organization's gaps and strengths (then suggest vectors to pursue accordingly)
- Assess an emerging technology to determine its likely development trajectory (especially commercialization)
- Help determine "so what?" as to how that emerging technology fits our organization's plans (road-mapping technologies and products)
- Help manage R&D processes – prioritize programs and projects better by providing empirical bases for decisions
- Inform IP-based strategic choices – help figure out "why?" a competitor is pursuing particular technologies and patenting strategies
- Improve other MOT decisions – technology insertion, national foresight, ...

REFERENCES

- [1] Katy Borner, Chaomei Chen, Kevin W. Boyack, "Visualizing Knowledge Domains," ARIST, Volume 37 09/30/2001
- [2] Michael Steinbach, George Karypis, Vipin Kumar, "A Comparison of Document Clustering Techniques" University of Minnesota, Technical Report #00-034 (2000). http://www.cs.umn.edu/tech_reports/
- [3] Gerard Salton, James Allan, Chris Buckley, Amit Singhal, "Automatic Analysis, Theme Generation and Summarization of Machine-Readable Texts," Science, Volume 264, 3 June 1994 1421-1426

- [4] Watts, Robert and Porter, Alan: Tracking the Evolution of Management of Technology (MOT), *International Association for Management of Technology (IAMOT) 2002 Conference*
- [5] Zhu, D. and Porter, A.L.: Automated Extraction and Visualization of Information for Technological Intelligence and Forecasting, *The 21st International Symposium on Forecasting, Pine Mountain, Georgia, June 17-20, 2001*.
- [6] Losiewicz, P., Oard, D.W., and Kostoff, R.N.: Textual data mining to support science and technology management, *Journal of Intelligent Information Systems* 15(2), 99-119 (2000).
- [7] Watts, Robert J., Porter, Alan L., Minsk, Brian, "Automated Text Mining Comparison of Japanese and USA Multi-Robot Research," *Data Mining 2004*, Malaga, Spain, September, 2004
- [8] Watts, Robert J., "Research Evolution in Robotics Fuzzy Control Technologies," *Association of Unmanned Vehicle Systems International (AUVSI) conference*, Baltimore, MD, July, 2003
- [9] Porter, Alan L., Watts, Robert J., Anderson, Timothy R., "Mining PICMET: 1997-2003 Papers Help You Track Management of Technology Developments," *Proceedings, Portland International Conference on Management of Engineering and Technology (PICMET)*, Portland, OR, USA, July, 2003
- [10] Watts, Robert J., Porter, Alan L., "R&D Cluster Quality Measures and Technology Maturity," *Technology Forecasting & Social Change* Vol. 70 pg 735-758 (2003)
- Huebscher, Markus, C., McCann, Julie A., "Simulation Model for Self-Adaptive Applications in Pervasive Computing," *2nd International Workshop on Self-Adaptive and Autonomic Computing Systems (SAACS 04)*, Zaragoza, Spain, Aug-Sep, 2004.
- Constantinescu, Z., "Towards an Autonomic Distributed Computing Environment," *14th International Workshop on Database and Expert Systems Applications*, 2003.
- Sterritt, R., Bustard, D. "Autonomic Computing – A Means of Achieving Dependability?", *10th IEEE International Conference and Workshop on the Engineering of Computer-based Systems*, 2003.

Appendix B

Technical Text Clustering Enhanced by Protected Natural Language Processing

Robert J. Watts
U.S. Army Tank Automotive Research,
Development & Engineering Center
bob.watts@us.army.mil

Alan L. Porter
Search Technology, Inc.
aporter@searchtech.com

Abstract

Recent discussions at ICDM and elsewhere address ways to enhance text processing. We have developed a practical mechanism to incorporate external controlled terms into natural language processing (NLP) of text segments. This “protected NLP” compares favorably to results from use of keywords, and also to regular NLP. We compare these for simple characterizations and for standardized factor (cluster) maps of IEEE conference content. A subject matter expert reviewed results to confirm considerable value added by this approach. Results point toward technological intelligence applications that profile emerging technologies based on conference content patterns.

1. Introduction

Science and technology (S&T) information resources continue to expand dramatically. Research publication and patent databases provide tremendous repositories that are complemented by web compilations. Together, these offer a magnificent source of technological intelligence to guide R&D management, intellectual property management, new product development, and related technology management processes.

Text mining tools can qualitatively enhance the utility of these information resources. Traditionally these have been searched to locate a handful of articles to be tracked down and read. Use of various indexing aids expedites these search and retrieval capabilities, but it also can offer far more. “Research profiling” can depict activities over an entire research domain [1, 2]. Trends in research activity and topical clustering can provide vital contextual information to help formulate effective research programs. Content analyses combined with information visualization can alert researchers, developers, and managers to intersecting domains, new methods, and gaps that offer best opportunities. Technical content text analyses can help answer technology managers’ “who, what, where, and when” questions.

Conference proceedings pose an intriguing text analysis challenge. Typically, an organization sends someone to attend, who may bring back personal, “internal” knowledge and contacts, as well as a conference CD. However, making the conference information more accessible to others remains hit or miss. We see high payoffs in changing this. Proceedings are increasingly available via multiple modes (e.g., IEEE Websites, databases such as IEE INSPEC and EI Compendex). Content can be treated at different levels: titles, keywords, class codes, abstracts, and full text. For our purposes, abstract records provide an excellent middle ground. From a technology management perspective, these are an underutilized technical intelligence resource. We thus set out to “mine” these texts [3].

A special challenge in mining conference proceedings from web sources is their lack of imposed indexing. For this experiment we therefore compare what can be gleaned from the “raw” web abstract records with that from databases. That is, we compile proceedings paper abstracts from a) conference websites, and also from b) databases that include these same conference papers. Elsewhere we compare technical intelligence based on single conferences, a sequence of conferences, and a more inclusive search on conferences plus journal articles [4]. In this paper we focus on text analyses, comparing results for two conferences from alternative data processing approaches.

We address three processes to tap technical content:

- Natural Language Processing (NLP) applied to paper titles and abstracts to extract noun phrases [5].
- Keywords, as provided by the INSPEC and/or EI Compendex databases; these include controlled keywords (i.e., index terms applied by the database upon importing the proceedings content) and author-provided keywords. These are also known as descriptors and identifiers, respectively. We also contrast with content characterization using the database class codes.
- “Protected” NLP, wherein we provide terms from an external source (here, the keywords from the databases) and instruct our NLP parser not to break any multiword keyword phrases.

2. Text Analyses for Content Mapping

The roots of our text mining lie in S&T analyses [c.f., 3, 6, 7, 8]. These key on the rich repositories of R&D publication and patent information – databases such as INSPEC, EI Compendex, Science Citation Index, Medline, and Derwent World Patent Index. Those make millions of articles and other reports available in semi-structured abstract records. We focus on co-term analyses, positing that more than expected co-occurrence across records can indicate relationship. We note that others use this information conversely to map records based on shared term usage – e.g., to map science [c.f., 9].

Abstract records provide far richer content for analysis than do citations alone. They are also far cleaner characterizations of the paper's focus than are full text records. As such, they are the prime resource for research domain profiling and for many competitive technical intelligence purposes. They also present special challenges. For instance, patent abstracts are not motivated to convey intent clearly, so introduction of external terminology to help index content can make a tremendous improvement [10, 11]. Our present use of sets of abstract records that appear as web resources and are also included in databases provides an ideal resource on which to compare analytical approaches.

Abstracts are typically available as semi-structured records. This aids straightforward analyses of certain fields. For instance, author information and publication date are usually cleanly set apart as fields. On the other hand, fields such as author affiliation may require parsing to extract organization and country (e.g., from “Departments of Industrial & Systems Engineering, and Public Policy, Georgia Institute of Technology, Atlanta, GA, USA”). Challenging text mining issues arise in treating titles, abstracts, and various forms of keywords. Merging such fields can enrich the accessible content markers.

We've found it fruitful to apply text mining to our own research domain for inherent interest and to help in the present literature review [5, 12]. We searched INSPEC on “text mining” and retrieved 342 abstracts for 2003-05 (on March 24) – it's an active domain! Text mining is a broadly disbursed specialty -- the papers are widely distributed among 119 conferences and 98 journals. The breakout by country is interesting. Despite INSPEC's inclination toward English, half of the eight countries with more than eight first-authored papers are Asian: US (90), China (49), Japan (27), Taiwan (17), Germany (16), Australia (14), South Korea (11) and UK (11). ICDM (2003) is the largest conference source represented.

The challenge of dealing with un-indexed text has become prominent as web text resources grow. This has been approached various ways. We are less interested in

those that require training, preferring approaches that can be applied to diverse topics, with prospects of semi-automation [3]. Text classification is one aspect. This can demand unique assignment (e.g., a given term can appear in only one category or cluster), but it need not. Our preferred topical representations allow for multiple assignment, as an important basis for linkage, but strive to minimize this. Various approaches “borrow” term structure from outside sources [c.f., 13].

Hybrid information retrieval and clustering schemes seek to enhance performance [c.f., 14, 15]. Various external term sources are becoming available (e.g., thesauri of database class codes, hierarchies such as the Medical Subject Headings (MeSH), chemical registries, and technical dictionaries [e.g., NASA aerospace dictionaries

[\[http://www.hq.nasa.gov/office/hqlibrary/pathfinders/aerodic.htm\]](http://www.hq.nasa.gov/office/hqlibrary/pathfinders/aerodic.htm)). Hotho and colleagues demonstrate utility in importing a core ontology to use in document preprocessing and clustering [16, 17]. Others adapt related strategies to aid in document classification [18].

3. Methods

In another paper [4], we describe processes to grab proceedings information from websites with the aid of WebQL agent software [<http://www.ql2.com/>] and how these can be exploited to generate useful technology management intelligence. We used a WebQL script to retrieve specific conference proceedings from the IEEE Conference Proceedings web site [http://www.computer.org/proceedings/proceed_a-h.htm]. To experiment with advanced text processing methods, we here analyze two sets of conference proceedings:

- The International Workshops on Database and Expert Systems Applications (“DESA”) for 2001-2004
- The Symposia on “Haptic” Interfaces for Virtual Environment and Teleoperator Systems

We selected DESA for inherent interest and also as a distinctly heterogeneous technical collection. Haptics offers a contrasting, relatively homogeneous conference.

We performed parallel data retrievals for the Haptics and DESA conference abstracts. We searched on EI Village in the EI Compendex and INSPEC databases. Those searches were constructed to retrieve similar topical coverage to that of the respective conferences, but of extended time spans and sources (various conferences and journals). These provide interesting evidence on how effectively mining conference proceedings serves to depict the state of the art in given research domains [4]. Here, we focus on records from the Haptics Workshops for 2002-04 and the DESA Workshops 2002-04 (but emphasize results for the 2004 Workshop).

We apply text mining software tailored for semi-structured data such as S&T abstract records. Three essentially equivalent versions of the software are *VantagePoint*, *Derwent Analytics* (tailored for Thomson Scientific data sources), and *Tech OASIS* (for US Government use) [//www.theVantagePoint.com]. *VantagePoint* serves to clean the data (e.g., apply fuzzy matching to consolidate term variations on import and during analyses), merge fields (e.g., controlled and uncontrolled keywords), and perform the essential analyses.

We draw keywords from an outside source (Dialog Link Format 8, free) to create a dictionary of domain-centric terms – an effective and reproducible way to produce an operational ontology [c.f., 19]. In other words, we search in databases via Dialog on terms that characterize the conference at issue (e.g., haptics) to collect keywords.

Prior to NLP parsing, the domain entities terms (i.e., the set of protected terms) are searched for and tagged in the abstracts' free text. The tagged terms/phrases are not parsed during NLP parsing. The resulting phrases field then includes the tagged domain entities plus NLP-parsed terms. We thus “protect” the domain entities term set to aid in describing the technical content.

We then performed this protected NLP import on the record sets gathered from the IEEE conference proceedings website. Uncontrolled vocabulary tends to be just that – tremendous variation resulting in low term frequencies. That, in turn, makes effective clustering hard. Instructing the NLP processor to protect established keywords for the research domain in question greatly improves this. For instance, for DESA 2004, unconstrained import yielded just 102 terms with frequency of occurrence of 3 or more. The “protected NLP” process to identify domain-relevant keywords (matching the list of these imported from a Dialog search in INSPEC) identified 335 terms occurring in 3 or more abstract records.

Our clustering approach builds on Principal Components Analysis (PCA). We augment this as “Principal Components Decomposition” with an optimization routine to maximize records inclusion in the fewest number of term factors (“clusters”) [20]. We apply PCA to term sets to generate co-occurrence based principal components. Since these reflect a basic form of factor analysis, we call the resulting term groupings “factors.” Because of the familiar use of “clusters,” we also use that terminology, although other clustering approaches can yield different forms (e.g., K-means, hierarchical clustering). This PCA approach allows terms to appear in multiple factors. [PCA is closely related to Latent Semantic Analyses as well.]

4. Experimental Results

To gain a feel for the differences between protected NLP phrases and keywords, Table 1 compares DESA occurrences for the top 10 keywords and the protected NLP occurrences of the same or related terms in 172 conference papers. Not all keywords (index terms) appear in the abstracts, and frequencies differ considerably. Conversely, the most commonly occurring NLP phrases here are: information (66 records), applications (49), approach (38), management (35), and web (35). Just as one would anticipate, indexing tends to consolidate – more records are associated with given informative terms. Our domain expert observed that even author-generated keywords are motivated differently (to draw in readers) than are abstracts (to describe what the researcher is doing).

Table 1. Term frequencies

Keyword	Key	NLP	NLP Variants
World wide web	53	0	web (35)
Database systems	51	3	database (28)
Data acquisition	42	0	acquisition (4)
Information analysis	34	0	analysis (13)
Mathematical models	33	0	modeling (14)
Information retrieval	34	5	
Semantics	32	11	
XML	31	8	
Algorithms	26	8	
Metadata	25	11	

Class codes consolidate variations even further. The top 5 INSPEC codes for the DESA papers are:

- data processing – 86
- computer applications – 77
- electronic equipment, etc. – 67
- computer software, data handling & appls. – 54
- database systems – 53

So one has a range of choices in depicting technical content from extremely granular to highly categorized: e.g., abstract words, NLP title phrases, protected NLP phrases, uncontrolled keywords, controlled keywords, and class codes.

Many informative breakouts are possible. Space precludes extensive exploration, but we note a few:

- “who + what” – profile leading authors in terms of their content emphases
- “what + when” – track the advent of term appearance and diffusion in this literature
- “where + what” – show which research institutions emphasize which topic clusters

We explored factoring based on the raw NLP abstract terms. This clustered only 66% of the 2004 DESA records. The “protected NLP” clearly yielded better record description than did raw NLP abstract phrases. Remaining comparisons consider “protected NLP”

analyses compared to use of the full array of keywords available in the database version.

Standardizing factor or cluster approaches offers tremendous advantages in enabling comparisons, semi-automation of generation, and consequent familiarization and popularization in technology management applications [21, 22]. We have devised a keyword-based standard PCA approach [23, 24]. This uses a repeatable factoring process to generate standard factor maps. Criteria include term cutoffs based on Zipf distribution characteristics and three cluster quality metrics:

- *Cohesion* – relatedness of abstract records within factor groups (higher cohesion is better)
- *Entropy* – overlap of records among factor groups (lower entropy is preferred)
- *F-measure* -- represents the maximum similarity (relatedness) between each factor and any of the other factors derived for a set of records. So, minimizing the maximum similarity provides a small F-measure (i.e., the desired result).

These factor quality metrics also offer potential interpretation in tracking S&T development. For instance, in time series of research publications relating to an emerging technology, reduced Cohesion may signal domain knowledge expansion; decreased Entropy may indicate spawning of divergent research branches; increased F-measure may imply convergence toward system integration.

The right number of terms to use in depicting a research domain is hard to specify. Our experiences suggest best results from excluding the few most common descriptors, then including some 100-250 terms, looking for a suitable “elbow” as threshold. Landauer made the case for somewhat more terms, on the order of 300 (ranging from 50-500) [25]. Our algorithm determines the number of factors to extract, from such term sets, based on the just indicated criteria.

Good cluster quality is defined as minimizing entropy and maximizing cohesiveness of the cluster groups [24]. Table 2 presents comparative results for the abstracts from the DESA and Haptics conferences.

Table 2. Cluster comparisons

DESA Workshop 2004	Protected NLP	Database Keywords
Entropy (ave. wtd/factor)	0.100	0.068
Cohesion (ave. wtd/factor)	0.059	0.062
F measure	1.36	0.84
# of Factors	22	13
% of Records Clustered	98%	97%
Haptics Symposia 2002-04		
Entropy (ave. wtd/factor)	0.093	0.022
Cohesion (ave. wtd/factor)	0.07	0.05
F measure	1.32	0.42
# of Factors	19	11

% of Records Clustered	97%	76%
------------------------	-----	-----

For both the relatively heterogeneous collection, Expert Systems (DESA), and the more homogeneous Haptics records, the keywords-based PCA analyses result in cleaner clustering (lower average entropy). Factor (cluster) group cohesion is quite comparable. The protected NLP abstract phrases result in a richer portrayal of conference content. For both conferences they yield more factors and equal or greater inclusion of records. So, in terms of parsimony, the keywords-based clustering is better. In terms of “richness,” an argument could be made in favor of the protected NLP approach. [Recall that “protection” borrows the controlled keywords from the database (INSPEC), then assures that these are found and not further split up (e.g., “expert systems” is not parsed into “expert” and “systems”)].

We ran numerous comparisons. Table 2 summarizes comparison of extracting abstract “protected NLP” phrases from the web-based data that mirror the controlled keywords (index terms, including “descriptors” and “identifiers”) brought in from outside. On the other side, we report a blend of three types of database keywords – controlled, main, and uncontrolled (author provided).

So, for Haptics, the EI Village (Database Keywords) map factors are lower entropy (as they also were for DESA). However, the PNLP-extracted (Protected) term factors are more cohesive. This is consistent with there being more factors extracted. Typically, more groups would mean smaller groups and greater cohesion. We favor a factoring approach that does this as we usually find that such smaller groups combine more conceptually related terms, making them easier to interpret.

A particularly interesting comparison shows in the different cluster maps (Figures 1 & 2). These reflect standard factor maps (PCA). They differ because the term sets (PNLP vs. keywords) differ. The maps themselves use multi-dimensional scaling to locate nodes (factors) in proximity to those with more association. Because this inherently distorts to fit into two dimensions, we use a path-erasing algorithm to indicate strength of association between any two factors. Location along the axes bears no inherent meaning.

Our domain expert observed: “I find the PNLP groupings more sensible, more complete. They seem to describe a body of knowledge that I would want to examine.” This seems to relate to the more complete coverage – 97% for PNLP vs. 76% for keyword-based clustering (Table 2).

However, our expert also noted that the associations among factors were more intuitive in the keyword-based clustering (Figure 2) than in PNLP (Figure 1). This may relate to “crisper” factors – note that the keyword-based factors are lower entropy (Table 2). They also group fewer records per factor (averaging 18.7 vs. 29.8 for the

PNLP factors). Obviously we cannot generalize based on a single expert user, but the target of low-entropy groupings to facilitate interpretability of relationships seems sensible.

5. Discussion

We have demonstrated that keywords provide a more effective representation of the technical content than do abstract NLP phrases. This holds in two quite different conference proceedings abstract sets – the heterogeneous DESA and the relatively homogenous Haptics IEEE conferences. This is no shock, and it confirms that database indexing definitely adds value.

More interesting are the comparisons between keywords alone and the combination of keywords with title or abstract NLP phrases. We introduce a novel method that provides both, even for an unindexed resource (abstracts from the IEEE website). It does so by identifying keywords from an external source (herein, the databases), then protecting these during NLP processing – PNLNLP. One could explore many other variations too – e.g., VantagePoint can extract Title NLP phrases and then merge this set with the “protected NLP” abstract phrases. One could then use this combination to profile record content and commonality.

PNLP has potential practical value in several regards. First, many commercial organizations do not subscribe to the databases on an unlimited use license. So, they can perform limited searches (e.g., in Dialog) to identify pertinent keywords for a topic, then incorporate these terms in depicting the content of a non-indexed information resource (e.g., abstracts retrieved from conference websites or CDs). Resulting analyses can offer both effective domain profiles and pointers to articles of prime interest (which might then be purchased via the databases).

Second, even where one has access to suitable keywords, it may be worthwhile to augment these with NLP phrases. As we demonstrate here, the combined term sets offer promise in providing a richer depiction of technical content. This could potentially be combined with further elaboration of the NLP-parsed text to add semantic structure to aid in interpretations [26].

Protected NLP also provides a mechanism to specify any terms, not just keywords, that one wants included in a full text analysis. For instance, one’s organization might have a lexicon that it applies to categorize a given technical area, so it would be highly valuable to tag articles according to these terms. This could directly enable “bucketing” of the records. It also could be blended with inductive clustering to yield multiple perspectives. We have previously demonstrated utility of a two-tier factoring process (Principal Components Decomposition [27]) that yields very helpful

categorization. For instance, breakouts on ceramic engine applications can be arrayed to easily see different application domains (e.g., ships, automotive) crossed with different topical emphases (e.g., fuel efficiency, pollution aspects).

In comparing different ways to analyze these conference proceedings, we also took note of the contrasts between conferences-only and broader coverage. Some quite intriguing patterns stand forth. For instance, in Haptics, we perceive three interest areas with relatively little overlap. Our INSPEC Haptics search included over 20% medical, whereas the IEEE Conference content that we analyzed included less than 10% in the medical domain. Ottawa University led with 39 publications, yet had none at the IEEE Haptics Conference.

Once the technically pertinent terms have been extracted for a set of documents (e.g., using PNLNLP), these can be profiled to understand the nature of the R&D and its progression over time [27]. Yoon and Park perform a full structural analysis of possible variation on each key dimension (morphological analysis) [28]. They track patent activity in a target domain to interpret innovative opportunities based on which energy sources, which alternative materials, and which process variants have and have not been tried. Text mining really does open new vistas to profile and understand research domain activity patterns and their implications.

6. References

- [1] A.L. Porter, A. Kongthon, J-C. Lu, Research profiling: Improving the literature review,” *Scientometrics* 53 (2002) 351-370.
- [2] K. Börner, C. Chen, K.W. Boyack, Visualizing knowledge domains, *Annual Review of Information Science and Technology* 37 (2003) 179-255.
- [3] A.L. Porter, S.W. Cunningham, *Tech Mining: Exploiting New Technologies for Competitive Advantage*, Wiley, NY, USA, 2005.
- [4] R.J. Watts, A.L. Porter, Mining conference proceedings for corporate technology knowledge management, in *Proc. Portland International Conference on Management of Engineering and Technology*, 2005.
- [5] R.J. Watts, A.L. Porter, S.W. Cunningham, D. Zhu, TOAS intelligence mining, an analysis of NLP and computational linguistics, in *Proc. First European Symp. on Principles of Data Mining and Knowledge Discovery*, 1997, Bergen, Norway, Springer-Verlag, NY, USA, pp. 323-334.
- [6] A.F.J. Van Raan (Ed.), *Handbook of Quantitative Studies of Science & Technology*, North Holland, Dordrecht, 1988; see also: <http://www.cwts.nl/>
- [7] R.N. Kostoff, E. Geisler, Strategic management and implementation of textual data mining in government organizations, *Technology Analysis & Strategic Management* 11 (1999) 493-525; see also: http://www.onr.navy.mil/sci_tech/special/354/technowatch/

- [8] A.L. Porter, Text mining for technology foresight," in T. Gordon, J. Glenn (Eds.), *Futures Research Methods, Millennium Project of the American Council for the United Nations* University, July 2003, http://www.acunu.org/millennium/FRM_v2.0
- [9] R. Klavans, K.W. Boyack, Identifying a better measure of relatedness for mapping science, *J Amer. Soc. For Inform. Sci. and Technol.*, to appear.
- [10] R.J. Watts, A.L. Porter, Requirements-based knowledge discovery for technology management, *Proc., Portland International Conference on Management of Engineering and Technology (PICMET)*, Portland, OR, USA, 2001.
- [11] R.J. Watts, A.L. Porter, Innovation forecasting using functional/capabilities analyses, *Internat. Symp. on Forecasting*, Edinburgh, 1998.
- [12] D. Zhu, A.L. Porter, S. Cunningham, J. Carlisle, A. Nayak, A process for mining science & technology documents databases, illustrated for the case of 'knowledge discovery and data mining,' *Ciencia da Informacao* 28 (1999) 1-8.
- [13] W.S. Lee, H. Wu, B.Liu, Semi-supervised text classification using partitioned EM, *Lecture Notes in Comput. Sci.*, vol. 2973 (2004), 482-493.
- [14] J. Kogan, C. Nicholas, V. Volkovich, Text mining with information-theoretic clustering, *Comput. Sci. Eng.* 5(6) (2003) 52-59.
- [15] E. Erosheva, S. Fienberg, J. Lafferty, Using mixed membership models for mapping knowledge domains, *Sackler Colloquium on Mapping Knowledge Domains*, Irvine, CA, USA, 2003; <http://vw.indiana.edu/sackler03/#Tentative>
- [16] A. Hotho, A. Maedche, S. Staab, Text clustering based on good aggregations, in *Proc. 2001 IEEE Internat. Conf. on Data Mining*, pp. 607-608.
- [17] A. Hotho, S. Staab, G. Stumme, Ontologies improve text document clustering, *Third IEEE Internat. Conf. on Data Mining*, pp. 541-544.
- [18] D. Barbara, C. Domeniconi, K. Ning, Mining relevant text from unlabelled documents, *Third IEEE Internat. Conf. on Data Mining*, 2003, pp. 489-492.
- [19] R. Prieto-Diaz, A faceted approach to building ontologies, *Proc., 2003 IEEE Internat. Conf. on Inform. Reuse and Integration*, pp. 458-465.
- [20] R.J. Watts, Knowledge discovery using the Tech OASIS: Meeting the information infrastructure needs *Proc., Portland International Conference on Management of Engineering and Technology (PICMET)*, Portland, OR, USA, 2001.
- [21] D. Zhu, A.L. Porter, Automated Extraction and Visualization of Information from Bibliographic Sources, [summary] *Proc., IJCAI-99 Workshop on Text Mining: Foundations, Techniques and Applications*, Stockholm, 1999.
- [22] D. Zhu, A.L. Porter, Automated extraction and visualization of information for technological intelligence and forecasting, *Technol. Forecast. and Social Change* 69 (2002) 495-506.
- [23] R.J. Watts, Knowledge discovery using the Tech OASIS: Meeting the information infrastructure needs, in: *Proc. Portland International Conference on Management of Engineering and Technology*, 2001.
- [24] R.J. Watts, A.L. Porter, R&D cluster quality measures and technology maturity, *Technological Forecasting and Social Change* 70 (2003) 735-758.
- [25] T. Landauer, D. Laham, From paragraph to to graph, *Sackler Colloquium on Mapping Knowledge Domains*, Irvine, CA, USA, 2003; <http://vw.indiana.edu/sackler03/#Tentative>
- [26] P. Sameer, K. Hacioglu, W. Ward, J.H. Martin, D. Jurafsky, Semantic role parsing: Adding semantic structure to unstructured text, *Third IEEE Internat. Conf. on Data Mining*, 2003, pp. 629-632.
- [27] R.J. Watts, A.L. Porter, Innovation forecasting, *Technol. Forecast. Soc. Change* 56 (1997) 25-47.
- [28] B. Yoon, Y. Park, A systematic approach for identifying technology opportunities: Keyword-based morphology analysis, *Technol. Forecast. And Social Change* 72 (2005) 145-160.

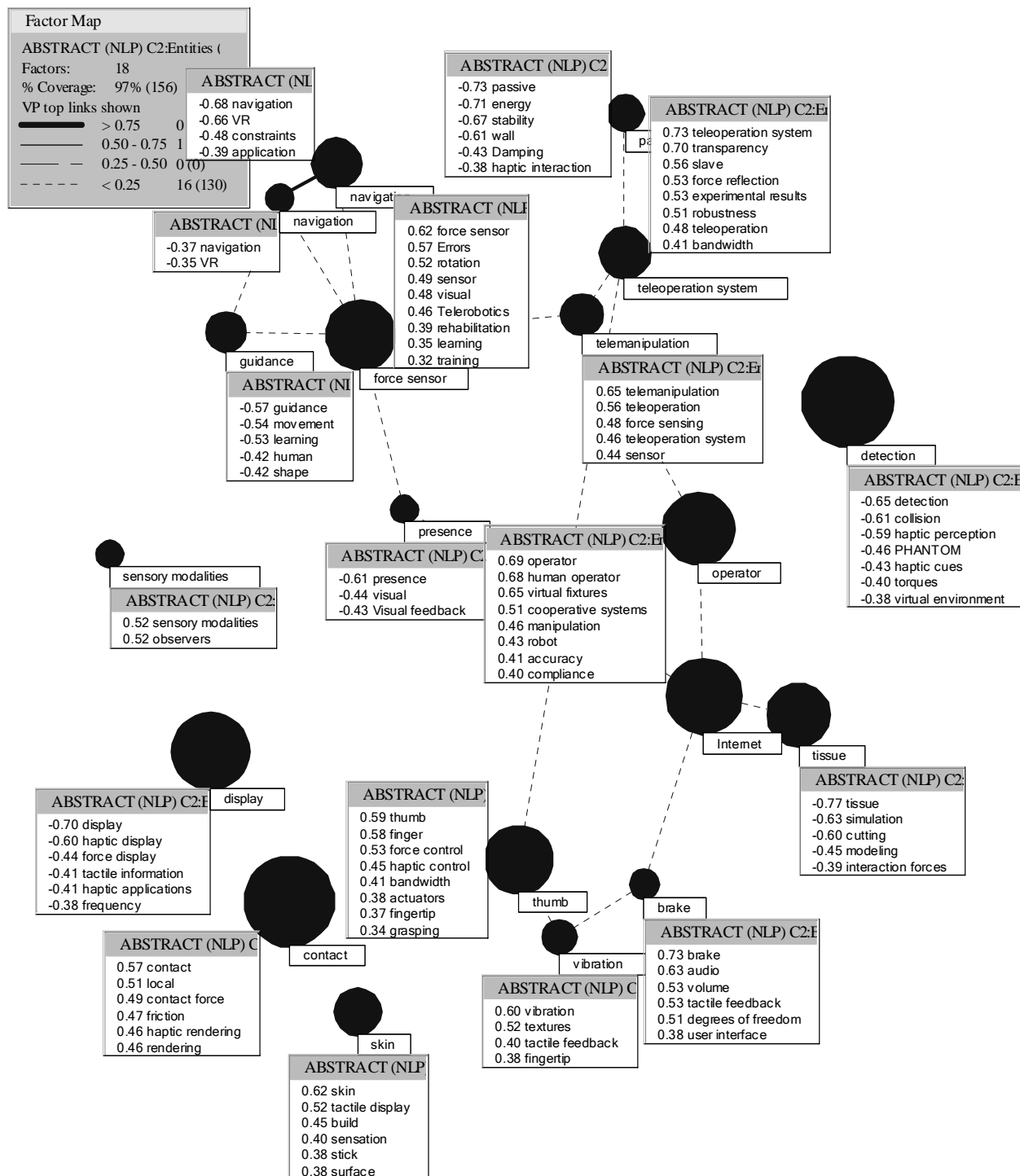


Figure 1. Haptics Factors based on protected NLP Abstract Phrases

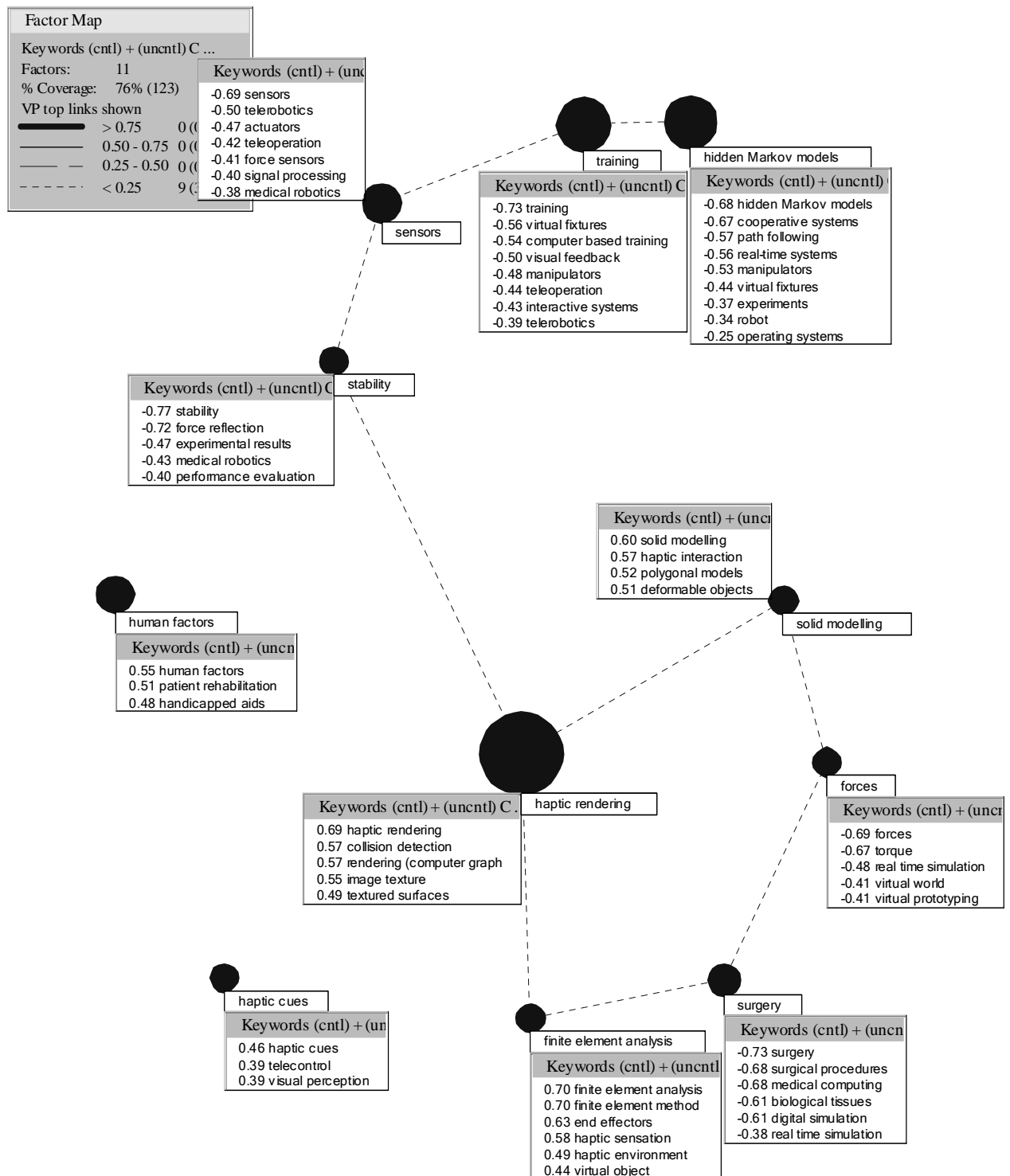


Figure 2. Haptics Factors based on INSPEC Keywords